Capsid Structure and DNA Packing in Jumbo Bacteriophages

by

Jianfei Hua

B.S., University of Science and Technology of China, 2004

M.S., University of Pittsburgh, 2010

Submitted to the Graduate Faculty of

the Kenneth P. Dietrich School of Arts & Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS & SCIENCES

This dissertation was presented

by

Jianfei Hua

It was defended on

March 30, 2016

and approved by

Andrew VanDemark, Ph.D., Assistant Professor, Department of Biological Sciences

Graham Hatfull, Ph.D., Professor, Department of Biological Sciences

Craig Peebles, Ph.D., Professor, Department of Biological Sciences

James Conway, Ph.D., Associate Professor, Department of Structural Biology

Dissertation Advisor: Roger Hendrix, Ph.D., Distinguished Professor, Department of Biological Sciences

Capsid Structure and DNA Packing in Jumbo Bacteriophages

Jianfei Hua, PhD

University of Pittsburgh, 2016

Jumbo phages, the phages with genome length larger than 200 Kbp, are extreme examples of how the capsid and genome coordinate in evolution. To learn the mechanism of capsid size change during evolutionary time, the capsid and DNA of several jumbo phages were characterized.

The capsid structure and protein of a T=25 *Sphingomonas paucimobilis* phage PAU and a T=28 *Escherichia coli* phage 121Q were studied in detail. The high resolution cryo-EM structures show that the major capsid proteins (MCPs) of both phages adopt the HK97 fold which is conserved in all solved tailed phage MCP structures. The capsids contain decoration proteins with unprecedented shape and location. A pentameric protein structure is attached on the inner surface of the pentamer in both capsids. The PAU capsid has arcs of density located on hexamers surrounding the pentons, which may bend the conformation of the subunit it interacts with to improve capsid stability. 121Q capsid contains dimeric density near the local 2-fold symmetry axes and knob-like density at the middle of the hexamer, which may participate in forming the capsid shell because the 121Q MCP leaves holes at the two locations. Both capsids contain a number of internal proteins whose roles are not clear.

The study on the jumbo phage DNA started with showing that partial modification of the cytosine in PAU and phage G DNA significantly slowed the phage DNA in electrophoresis. A new technique was developed to quantify the effect of base modifications in large DNA in

electrophoresis, with which reliable measurements of the chromosome size of our jumbo phage collection were made. Jumbo phages have larger terminal redundancy and lower DNA packing density compared to small and mid-sized phages. The result on the DNA packing density in different sizes of phages reveals a negative correlation between the capsid size and the DNA packing density. We explain this relationship by a model based on the strength limit between the capsomers of the capsid shell.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

My graduate career is a long path. I received so much help from so many people. First and foremost I would like to thank my advisor, Roger Hendrix, for his persistent and generous support and guidance for so many years. I would like to thank Bob for supervising my projects and carrying me though all the technical difficulties in my experiments. I would like to thank James for providing continued training and collaboration. I would like to thank my committee members, James, Craig, Andy and Gram for all the help and insight during the meetings.

I would like to thank our collaborators from James's lab, Alexis and Katarina, from their great work on the cryo-EM structure of jumbo phages. Thank you Susan for giving all the advice during our lab meeting. Thank you to past and current members in the Hendrix's lab, Aletheia, Bonnie, Brian, Dan, Dan-Ju, Jacob, Josh, Patricia, Rob, Welkin and Xiaoxian, for the help, support and friendship. And my family, for their love and understanding. It is really great to close this chapter of my life.

1.0 INTRODUCTION

1.1 BACTERIOPHAGES

1.1.1 Why study phages?

My projects focus on the lesser known jumbo bacteriophages (Hendrix 2009). Bacteriophages, or phages, are viruses infecting Eubacteria, and some of the viruses that infect Archaea are related to tailed dsDNA phages both morphologically and (apparently) evolutionarily. Bacteriophages are the most ubiquitous and abundant form of life in the biosphere. Phages have been an important subject of biological research for a century since the independent discovery by Frederick W. Twort (Twort 1915) and Felix d'Herelle (d'Herelle 1917). In the 1960s and 70s, studies on phages led to some of the most exciting discoveries and innovations in biological research, such as the (Hershey and Chase 1952), the genetic codon (Nirenberg and Matthaei 1961) and the development of phage enzymes and plasmids as genetic tools (Lederberg 1952; Sugimoto, Okazaki et al. 1968). In recent years, phages have attracted increasing attention of scientists from different fields. One reason for that, in my opinion, is the development of new technologies applicable in phage research, such as massive and inexpensive sequencing technologies that make comparative genomics in phage research practical (Ronaghi, Karamohamed et al. 1996; Brenner, Johnson et al. 2000), and high-resolution cryo-EM

technologies that provide structure of phage particles in great details (Adrian, Dubochet et al. 1984; Conway, Cheng et al. 1997; van Heel, Gowen et al. 2000). Another reason is about our renewed understanding of the ecological importance of phages to environments, first shown by the surprising ubiquity and abundance of phages in the sea (Suttle 2005). And lastly, phage therapy, the therapeutic and industrial application of phages, is a promising alternative to antibiotics to fight against the growing problem of antimicrobial resistance (Keen 2012).

1.1.2 Classification of phages

Phages are diverse in terms of morphology and nucleic acid. However, most phages currently known are, and most of our understanding about the nature of phages and the environmental impact of phages is from, members of the tailed phages of the order *Caudovirales* under the current classification by the International Committee on Taxonomy of Viruses (ICTV) (Ackermann 2001; Fauquet and Fargette 2005). The tailed phage virion contains an icosahedrally symmetric protein capsid enclosing a linear dsDNA genome, connected to a tail functioning as a host-infecting apparatus through a portal complex at one of the 12 vertices. Three families are currently recognized by the ICTV based on the morphology of the tail, including *Myoviridae* (long, contractile tail), *Siphoviridae* (long, noncontractile tail) and *Podoviridae* (short tail). However, many researches embrace the idea of "All the world's a phage", arguing that the horizontal gene exchange between genomes of the dsDNA phages are so extensive that all the genomes of the dsDNA phages are linked together, forming a large common genetic pool (Hendrix, Smith et al. 1999). In this scenario, boundaries between the dsDNA phage species are not well-defined, and isolation of different phage strains are just biased sampling of the part of the pool accessible to us.

1.1.3 Two lifestyles

Regardless of tail morphology, tailed phages may live two lifestyles: the lytic life cycle and the lysogenic life cycle. To infect a host cell, a tailed phage attaches to the host surface by interaction between the tail fiber and some specific receptor on the cell surface. The phage DNA is transferred through the portal and the central channel of the tail and injected into the host cell. In the lytic cycle, the host metabolism is taken over by the phage to reproduce phage parts which eventually assemble into numerous new progeny phages. Soon after, the new phages are released by attacking and breaking the host cell with endolysin. However, a temperate phage may choose the lysogenic path when its DNA enters the host, which does not result in immediate lysing. The phage DNA becomes inactive either as a form of a plasmid or an integrated part of the host chromosome, and grants the host immunity to similar phages. While the host cell continues to grow and reproduce, the dormant phage DNA, or the prophage, replicates and segregates into all the daughter cells. The prophage may become active and enters the lytic cycle when the host condition deteriorates.

1.2 THE CAPSID OF TAILED PHAGES

1.2.1 Quasi-equivalent virus capsids

In 1956, Crick and Watson pointed out that a virus capsid could not be constructed by single copies of many different proteins because the genome enclosed in such a capsid would not be long enough to encode all these proteins (Crick and Watson 1956). Instead, they proposed that

the capsid structure could be based on one of the cubic symmetries, so multiple copies of the same protein in identical chemical environments could form the capsid. The icosahedron was later found to be the predominant shape in phage capsids, and two other shapes, the prolate icosahedron and the helical shape, occur much less frequently. The icosahedron surface contains 12 pentagonal vertices and 20 equilateral triangles constructed by 60 asymmetric units which are the identical subunits in the capsid from Crick and Watson's perspective.

Later in 1962, Caspar and Klug proposed the physical principles about how larger icosahedral capsids could be built by more than 60 copies of chemically identical subunits (Caspar and Klug 1962) (Figure 1). A flat hexagonal will be able to curve into an icosahedral shell if 12 properly selected hexagons are replaced by pentagons. For the simplest capsid with only 60 subunits, the environment and bonding of every subunit is the same. For larger capsids with more subunits, each asymmetric unit contains several "quasi-equivalent" subunits which are chemically identical subunits with slightly different environments and bonding. Only a series of selected numbers, called triangulation numbers (T-numbers) of subunits, are able to form asymmetric units in capsids constructed with the "quasi-equivalence" principle. The T-number, which determines the geometric figure of the capsid, is restricted to

$$T = h^2 + h*k + k^2$$

where h and k are the non-negative integers that describe the coordinates of a neighboring pentagon to the origin in the lattice. The total number of subunits in an isometric icosahedral capsid is 60*T with T types of slightly different environments.

Figure 1 Construction of icosahedral capsids by a hexagonal lattice.

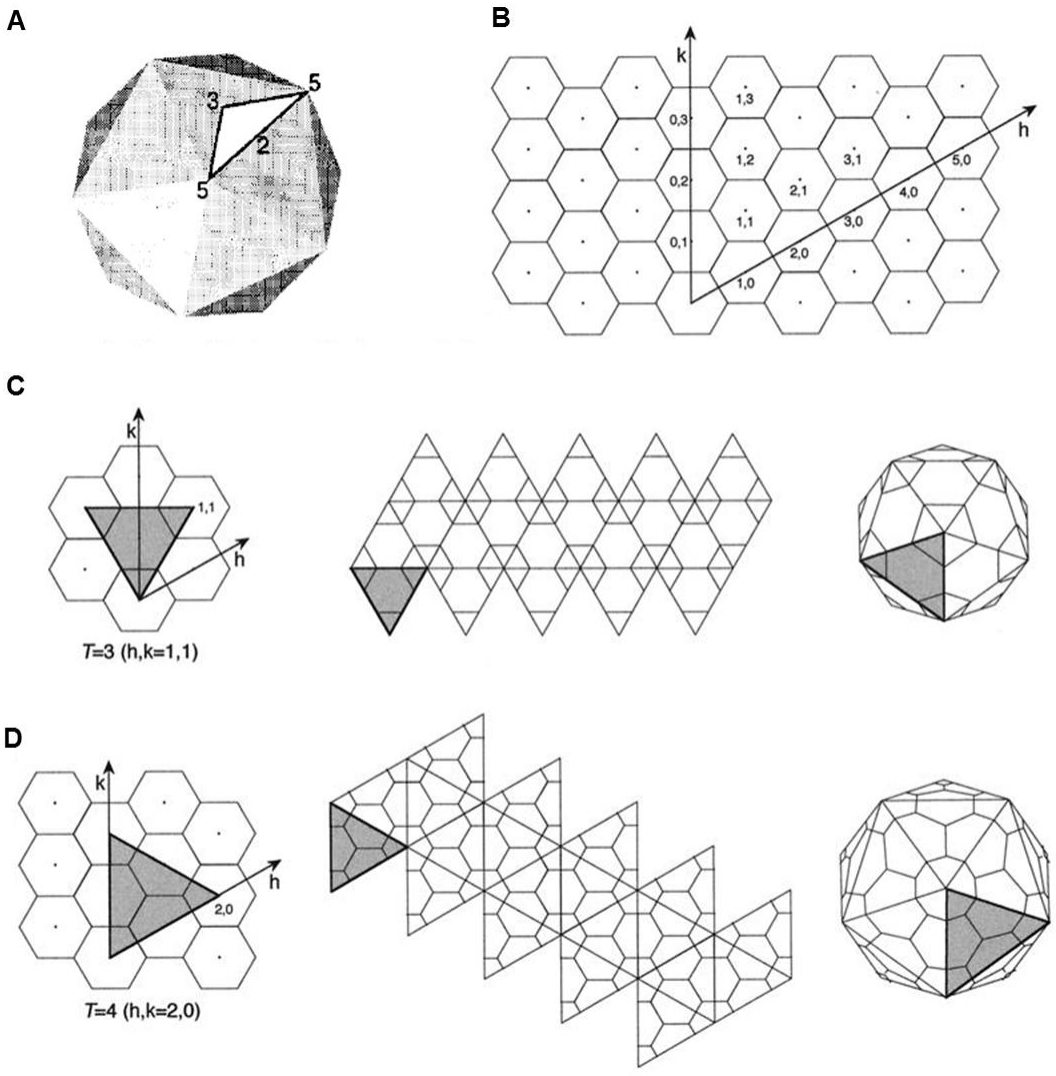
This figure is adapted from (Baker, Olson et al. 1999) with permission.

A. An icosahedron with symmetry axes marked.

B. A flat hexagonal grid. Numbers near the center of a hexagon are the h and k values of the capsid when this hexagon and the origin are converted to pentagons.

C. Construction of a T=3 capsid.

D. Construction of a T=4 capsid.



1.2.2 Capsid proteins

In most tailed phages, the quasi-equivalent subunit is formed by a single protein called the major capsid protein (MCP) or coat protein. In the T=7 phage HK97, the MCP gp5 is the only phage protein absolutely required to form the icosahedral capsid (Duda, Martincic et al. 1995). The MCP is highly conserved in tailed phages and in some icosahedral human viruses such as herpes simplex viruses, as we observe the same protein fold first discovered in HK97 gp5 in all solved MCP structures of these viruses despite very little sequence similarity (Hendrix 1999). Built by the conserved MCP, the dimension of an icosahedral capsid is almost entirely determined by the number of participating MCP copies which can be conveniently described by the T-number.

One of the twelve pentons is replaced by the portal, a dodecameric hollow protein ring which connects the capsid and the tail. Phage DNA is transferred through the channel in the middle of the portal during phage assembly and infection. In phage T4 and maybe all other tailed phages, the portal also plays important roles in initiating capsid assembly on host cell membrane (Kato and Baschong 1997).

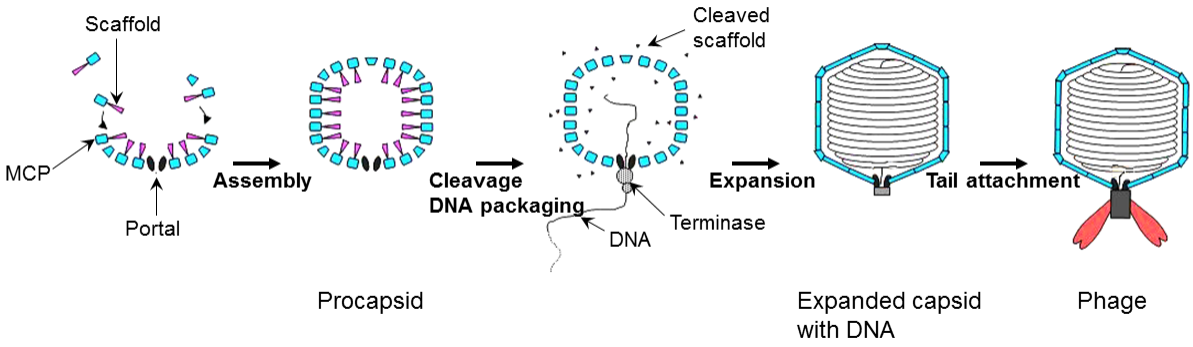
In addition to the MCP and the portal, some phage capsids contain accessory proteins on the capsid shell, called decoration proteins, and/or inside the capsid, called internal proteins. Decoration proteins, such as gpD in phage λ, make the capsid more stable, but are usually dispensable in laboratory conditions (Lander, Evilevitch et al. 2008). Internal proteins may coordinate the MCP during capsid assembly (Kurtz and Champe 1977) or protect the phage DNA during infection (Bachrach and Benchetrit 1974).

1.2.3 Capsid assembly

The assembly pathway of tailed phage capsids is a well-defined process in which each step only proceeds when all the previous steps are finished (Casjens and Hendrix 1988), reducing undesirable production of dead-end parts lacking necessary components. The capsid assembly pathways in tailed phages share considerable similarities, in spite of high variations in the structure and protein of the capsid and in details of individual assembly (Casjens and King 1975; Casjens and Hendrix 1988). We believe that the assembly starts around the portal (Figure 2). With assistance from the scaffolding protein, the MCP polymerizes into a round particle called procapsid which contains the protease inside (the protease is not drawn in the figure). In many phages, the N-terminus of the MCP, the scaffolding protein and the protease itself are proteolytically processed by the protease and the cleaved short peptides exit the capsid. The round procapsid then expands into an angular, bigger and more stable capsid, while the phage DNA is packaged into the capsid through the portal by the terminase. In some phages, decoration capsid proteins are attached onto the surface of the expanded capsid. The tail is either added to the portal of the DNA-filled capsid as a preassembled part in phages with long tails, or assembled on top of the portal by sequentially adding tail proteins (Casjens and Hendrix 1988).

Figure 2 General capsid assembly in tailed phages.

This figure is adapted from Dr. Weng Jiang's web figure with permission.



1.3 JUMBO PHAGE

While dozens of phages are among the simplest and best-understood organisms, those tailed phages with genome larger than 200kb, a.k.a the jumbo phages (Hendrix 2009), represent a very poorly understood corner of phage biology. Studies on a few examples, such as the Pseudomonas phage phiKZ and vibriophage KVP40, show some common characteristics in these tiny giants (Krylov, Smirnova et al. 1978; Mesyanzhinov, Robben et al. 2002; Miller, Heidelberg et al. 2003; Fokine, Kostyuchenko et al. 2005). Most jumbo phages have a large isometric capsid with maximum diameter exceeding 120 nm and T-number greater than 16. Their tails are all contractile, similar to the contractile tail in mid-sized phages like T4. Decoration capsid proteins and internal capsid proteins are commonly found in jumbo phages. In phiKZ, the internal proteins form a rod-shaped protein core inside the capsid (Wu, Thomas et al. 2012).

The large genomes of jumbo phages often encode more than 300 genes. Less than half of the genome contains conserved genes encoding structural proteins that may be matched with capsid and tail proteins in T4-like phages by sequence alignment (Hendrix 2009). Organization of these structural genes in jumbo phages is largely similar to mid-sized phages, but more often other less conserved genes are inserted between the structural genes. What really distinguishes the genomes of jumbo phages from those of mid-sized phages is that half or more of the genome is filled with genes which either do not match any other gene or match only putative genes in the GeneBank database. Genes involved in DNA replication and nucleoside metabolism are also found in the genomes of jumbo phages, with the exception of phiKZ (Hendrix 2009). However, this property is not unique in jumbo phages because it has also been noted in the genomes of mid-size phages like T4.

1.4 EVOLUTION OF CAPSID SIZE

It is widely accepted that all the currently documented tailed phages are evolved from the same ancestor, evidenced by conserved features in the remarkably diverse phage population, including the HK97 fold, the quasi-equivalent capsid and the shared capsid assembly pathway. Therefore, the capsids that we find today with a wide range of sizes may have evolved from the ancestor capsid of a certain size, which leads to the question of which is the preferred evolutionary trend in term of capsid size, to be smaller or to be bigger. We believe that increasing the capsid size is the more likely choice for tailed phages because the size of phage genome is restricted by the size of capsid and only a number of discrete sizes are allowed by the quasi-equivalence principle (Hendrix 2009). This hypothesis can be explained in phages with headful DNA packaging during which DNA packaging by the terminase continues until capsid is full (Black 1989; Tavares, Lurz et al. 1996).

Decreasing the capsid size likely ends with phages that are unviable or unfavorable in selection. The genome of a small phage like λ is usually compact and most genes are important structural genes. However, the genome of a mid-sized phage contains non-essential genes that are located all over the genome and mixed with the important genes. When the mid-sized phage produces a capsid with a smaller T-number, the genome must shrink its size to fit into the smaller capsid in a way that all the important genes need to be preserved to make the smaller new phage viable. It can be accomplished by deletion of many non-essential genes that are scattered in the genome in a single evolution event, which is very unlikely to happen considering how organized and complicated such a deletion would need to be. It can also be accomplished by accumulation of multiple separated deletions over a long period of time. The size of the capsid does not decrease until enough deletions have been accumulated. However, this process is counter-

selected, because the disadvantage of losing a non-essential gene that increases fitness is immediate but the benefit of saving materials in building a smaller capsid can only be received after the whole process is completed (Hendrix 2009).

A phage with mutation that suddenly expands the capsid size is more likely to survive. The large capsid immediately provides the space for the genome to acquire new genes. The newly acquired gene may not be able to initially offset the additional cost of assembling a larger phage, but it provides a starting point for achieving novel beneficial function while retaining the important genes from the original genome. Once enough new genes are included in the genome, the option to go back to the original size is no longer available to the phage as discussed above.

1.5 THESIS PLAN

Our long-term goal is to understand how tailed phages maintain and change their capsid size. And of course such a fundamental problem requires more than one PhD thesis to solve. I approached this problem by two directions: one is about the capsid structure and DNA of the less studied jumbo phages with novel T-numbers, and the other is about the procapsid structure of phage P1 which naturally produces three differently-sized capsids. This document is mainly a summary of the results of jumbo phages.

The document consists three main chapters. The first and second chapters discuss the data on the structure and protein of two jumbo phages, PAU and 121Q. Both of them are jumbo phages with novel T-numbers. The capsid structure of PAU and 121Q was solved by high-resolution cryo-EM microscopy and 3D reconstruction. The capsid protein was identified by N-terminal sequencing and mass spectrometry. The fold of the MCP was studied by fitting the

HK97 MCP fold into the cryo-EM density map. The location of the capsid protein on the capsid was determined by combining the cryo-EM structure and the biochemistry of the capsid.

The other chapter covers the result on the DNA size of jumbo phages. We started with verifying some previous estimates of the chromosome length of phages PAU and G which were much higher than the genome size. A new technique was developed to estimate the effect of base modification in large DNA in electrophoresis, with which reliable measurement of the chromosome sizes of our jumbo phage collection were made. The measured DNA packing density in different sizes of phages reveals a negative correlation between the size of the capsid and the DNA packing density. A model based the tension of the capsid shell is proposed to explain this negative correlation.

The whole project is not a well-planned and hypothesis driven one in an established field, but rather an expedition to an unfamiliar area of phage biology.

2.0 STRUCTURE AND PROTEIN COMPOSITION OF *SPHINGOMONAS PAUCIMOBILIS* PHAGE PAU VIRIONS

2.1 INTRODUCTION

2.1.1 *Sphingomonas paucimobilis*

Sphingomonas paucimobilis is the host of bacteriophage PAU, the subject of this chapter. *S. paucimobilis* is a non-fermenting Gram-negative bacillus, one of the best-known species of the genus *Sphingomonas*. The rod-shaped cell of *S. paucimobilis* is about 0.7 X 1.4 um in size, with a single polar flagellum and low mobility (Brenner 2005). *S. paucimobilis* is an oligotrophic bacterium, meaning that it that can grow in low nutrient environments (Tada and Inoue 2000). It has been identified in a wide range of environments such as soil, natural water, hospital and various water distribution systems (Ryan and Adley 2010). It can form biofilms in environments with continuous water, commonly seen as the slippery orange stain in bathrooms and kitchens. S. paucimobilis is capable of breaking down and utilizing a wide variety of dimeric lignin compounds (Masai, Katayama et al. 1999).

Although it is low in virulence and not considered a major pathogen, *S. paucimobilis* is an opportunistic pathogen widely found in clinical environments. *S. paucimobilis* is able to pass through 0.2 mm filters that are commonly used for sterilization of medical solutions and purified

water (Glupczynski, Hansen et al. 1984). The abilities to survive in low nutrient environments and to pass the 0.2 mm filters lead to many cases of *S. paucimobilis* infections in patients whose immune system is weakened.

2.1.2 Phage PAU

Bacteriophage PAU, the sole phage of *S. paucimobilis* currently known, was first isolated from diseased silkworms (*Bombyx mori* L.) by Ackermann and his colleagues in 1992 in India (Ackermann, Auclair et al. 1994). Homogenized silkworms were filtered with filter paper and centrifuged, and then examined by negative stain electron microscopy (EM). Besides the large amount of bacteria, various types of phages were found in the supernatant. They observed numerous phage particles with giant heads and relatively short tails, 156 nm in diameter and 119 nm in length respectively, named "X particles". A number of hosts and growth conditions were tried but no more X particles were produced. Instead, another large phage with the host identified as *S. paucimobilis* was isolated, named "PAU" after its host. The PAU virion had a head 112 nm in diameter and a tail 160 nm long, and an unusually large-sized DNA chromosome, measured to be about 412 kb by these authors using pulsed field electrophoresis. The discovery of PAU is a fine example of the ubiquity and diversity of bacteriophages.

In previous work I reconstructed a low resolution cryo-EM model of phage PAU which showed that the PAU capsid has a triangulation number of 25, so far unique among known phage structures, and a distinctive arrangement of minor capsid proteins both outside and inside the capsid (Hua 2010). Here I present more detailed studies on the structure and protein composition of the PAU capsid. We analyzed the three-dimensional (3D) structure of PAU mature capsids

with or without DNA by electron cryo microscopy (cryo-EM). The T=25 PAU capsid has minor capsid proteins arranged beneath the pentamer on the inner surface and also on the outer surface of the hexamers immediately surrounding the pentamers. At least 7 internal capsid proteins were identified, forming a loosely interacting structure in the mature empty capsid but not the same as the "inner-body" (Wu, Thomas et al. 2012) described in some capsids filled with DNA. The dissociation of the minor capsid proteins on the outer surface and the internal proteins under certain conditions enabled us to identify the genes that encode the above proteins. These show little sequence similarity to existing genes in GeneBank. The triangulation number and the organization of minor capsid proteins are unique to PAU among phages previously characterized in these regards, and low sequence similarity between the genomes of PAU and other phages, suggests that PAU may represent a distinct branch in tailed phage evolution.

2.2 PURIFICATION OF PAU PARTICLES

2.2.1 Morphology of Phage PAU

PAU has an isometric icosahedral capsid, a contractile tail connected by a narrow neck, and a narrow but long "baseplate" with several tail fibers attached (Figure 3). The PAU capsid diameter is 130 nm along the 5-fold symmetry axis and 116 nm along the 3-fold symmetry axis. The 170 nm whole tail including the neck and the baseplate results in a 300 nm whole phage from the distal vertex to the baseplate.

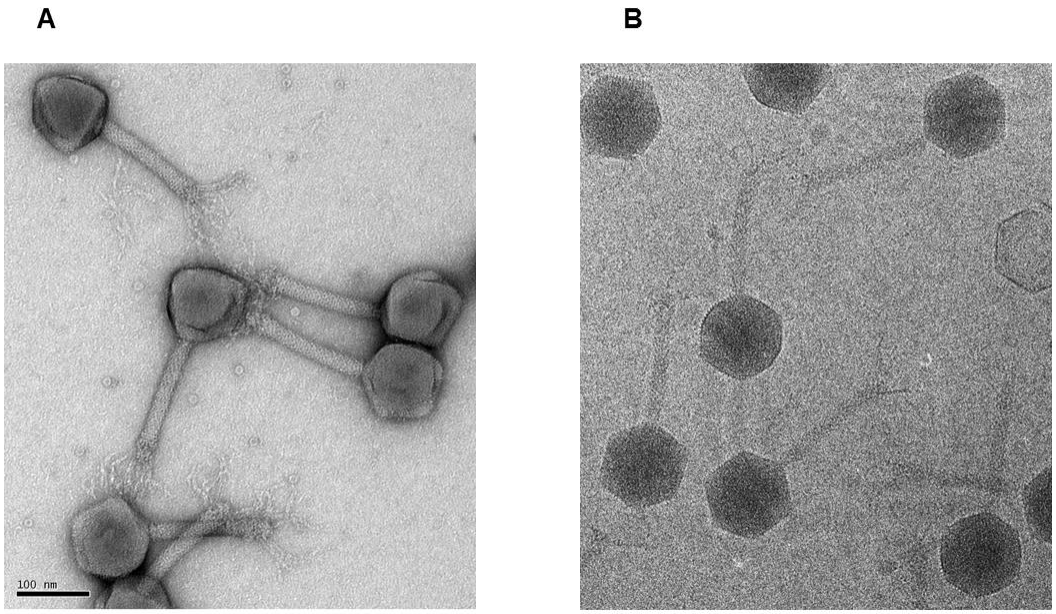
The negative EM micrograph shows that the tail fiber and baseplate of PAU appear different from those of a typical T4-like phage with a contractile tail. In the prolate phage T4, the

proximal and distal half-fibers link together with an angle about 20 °, forming a long, straight and narrow fiber as long as 140 nm (Leiman, Arisaka et al. 2010). The tail fiber of PAU is much shorter, measured at about 65 nm in the micrograph (Figure 3A). The trace of the tail fiber is curly and floppy, with a wider distal end that extends far away from the tail. The baseplate of T4 is notably wider than the tail sheath and can be easily recognized in electron micrographs; tail fibers are typically attached on the side of the baseplate and spread around the two sides of the baseplate in electron micrographs. However, the baseplate in PAU appears as narrow as the tail sheath and has an extra narrower structure, about 15 nm wide and 17 nm long, attached to the bottom of the baseplate (Figure 3A). The proximal end of the PAU tail fiber seems to interact with the bottom of the baseplate, so that most tail fibers extend far away from the capsid end in the electron micrograph.

Figure 3 EM pictures of PAU phages

A. A negative stain EM image of phage PAU. Both images share the same scale.

B. A cryo-EM image with phages and an empty capsid of PAU.



2.2.2 Growth of PAU

PAU was easy to grow compared to other jumbo phages I worked with. When mixed with fresh overnight *Sphingomonas paucimobilis* cells in LB with 0.25% agarose, the phage formed big, clear plaques with a defined edge between the plaque and bacterial lawn. Plate stock was an efficient method to produce high-titer PAU stock. On average, we obtained about 100 ml 10^{11} plaque forming units (pfus) per ml stock from plates containing 1 L LB with 1.5% agar. However, the color of the stock was orange because of cell debris with an orange pigment characteristic of *S. paucimobilis*.

2.2.3 Purification of PAU particles

It was necessary to obtain purified PAU empty capsids for further investigation of the PAU capsid. Such a capsid without a tail and DNA would allow us to accurately identify capsid proteins, and obtain structures of the minor capsid proteins inside the capsid without interference from the highly compressed DNA. The orange color in the PAU plate stock indicated the presence of host debris. Running the concentrated PAU stock through sucrose gradients generated a narrow empty capsid band and a wide phage band but some cell membrane still remained in both bands, indicated by the color and later confirmed by EM. Equilibrium CsCl gradient at 1.5 g/ml produced a clean white phage band near the middle of the gradient and a mixture of phage parts and host debris. The separation of PAU phage by equilibrium CsCl gradient suggested that PAU was stable in high concentrations of CsCl and CsCl gradients could effectively separate PAU particles from cell debris.

The CsCl step gradients with 30% sucrose, 1.4 g/ml CsCl and 1.6 g/ml CsCl provided an excellent separation of both phages and empty capsids from host components (Figure 4A). The light-weight orange host debris stayed in the middle of the 30% sucrose layer. A white narrow band at the interface between the 30% sucrose and 1.4 g/ml CsCl layers contained the empty protein capsids with a density close to 1.4 g/ml. A strong band at the interface of the 1.4 and 1.6 g/ml CsCl layers was phages with a density at about 1.5 g/ml. The presence of an abundant protein at about 50 kDa in the SDS gel, the expected size of a major capsid protein (MCP), in both sucrose bands confirmed that the sucrose bands were formed by PAU particles (Figure 4B). A number of bands that were relatively stronger in the empty capsid lane than in the phage lane were probably formed by other capsid proteins.

The empty capsid band was dialyzed and then further purified by running through 15~45% sucrose gradients. Three bands were formed after the centrifugation and EM examination showed that the middle major band contained clean empty capsids (Figure 4C & D). The sample was very pure, almost free from any host membrane. Empty phages were occasionally found when searching the EM grid. This was the capsid sample I used in the following studies on the PAU capsid.

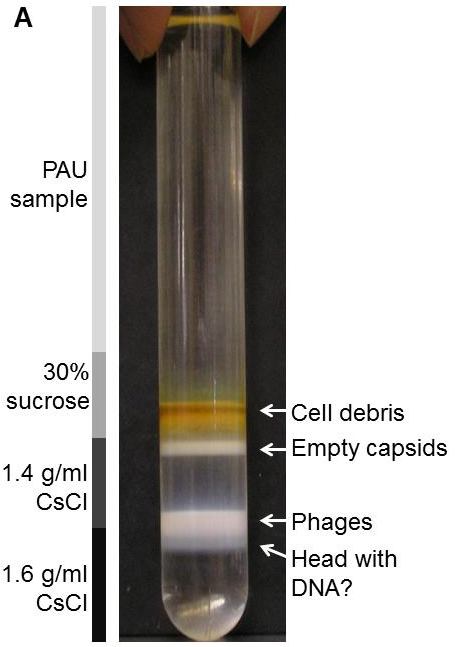
Figure 4 Purification of PAU empty capsids

A. A picture of the bands of the concentrated PAU plate stock analyzed in a CsCl step gradient. From top to bottom, the initial step gradient in a thin-wall SW41 tube contained 6 ml PAU, 2 ml 30% sucrose, 2 ml 1.4 g/ml CsCl and 2 ml 1.6 g/ml CsCl, as illustrated on the left. Four bands were observed from top to bottom containing: host debris, empty capsids, phages and presumably DNA-filled heads. The last band was too close to the phage band so it was not collected.

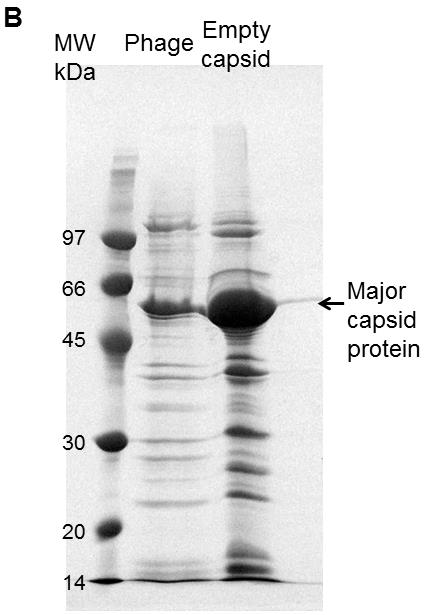
B. SDS gel of the PAU phage and the empty capsid bands. The most significant band at 50 kDa in both lanes is the major capsid protein. A number of bands that were relatively stronger in the empty capsid lane than in the phage lane were taken to be potential capsid proteins.

C. A schematic of the band pattern of the dialyzed empty capsid band in A running through 15%-45% continuous sucrose gradients. Three bands appeared and the middle empty capsid band was the strongest.

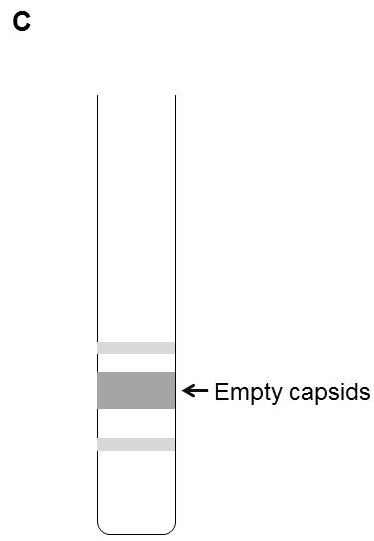
D. A negative stain EM image of the middle band in Figure 1C. The picture shows that the sample contained pure empty particles with a diameter ~ 1300 Å.



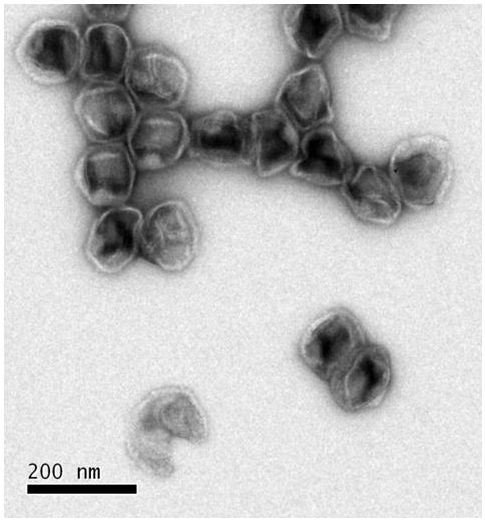
CsCl step gradients



SDS gel of CsCl step gradients bands



Schematics of sucrose gradients



Negative stain EM of PAU capsid

2.3 PROTEINS OF THE PAU CAPSID

2.3.1 Identification and quantification of PAU capsid proteins

The PAU capsid consists of more than a single major capsid protein, for the SDS gel of the cesium-purified PAU capsids displays many bands in addition to the major band (Figure 4B). We were interested in the function of PAU capsid proteins, in particular if any of them were related to the extra morphological features visible on the existing EM structure. The PAU capsid proteins were identified by Dr. John Hempel using Edman degradation (Edman 1950), which also provided the site of proteolytic processing by the protease at the N-terminus. The tail protein bands were easily identified from this analysis based on the reduction of the amount relative to the major capsid protein in the empty capsid sample compared with the phage sample.

The fact that a T=25 icosahedral capsid contains 1495 copies of the major capsid protein (MCP) enables estimates of the copy number per capsid of other proteins if their quantity relative to the MCP is known. Gel quantification is notoriously inaccurate in dealing with bands that have large differences in intensity, as band quantification is no longer linear to the actual quantity of the protein when two bands have significant difference in band intensity. Therefore, a series of dilutions of the PAU empty capsid sample were electrophoresed on the same SDS gel (Figure 5). Comparing bands with close intensity provided more accurate quantification of the relative intensity (Table 1).

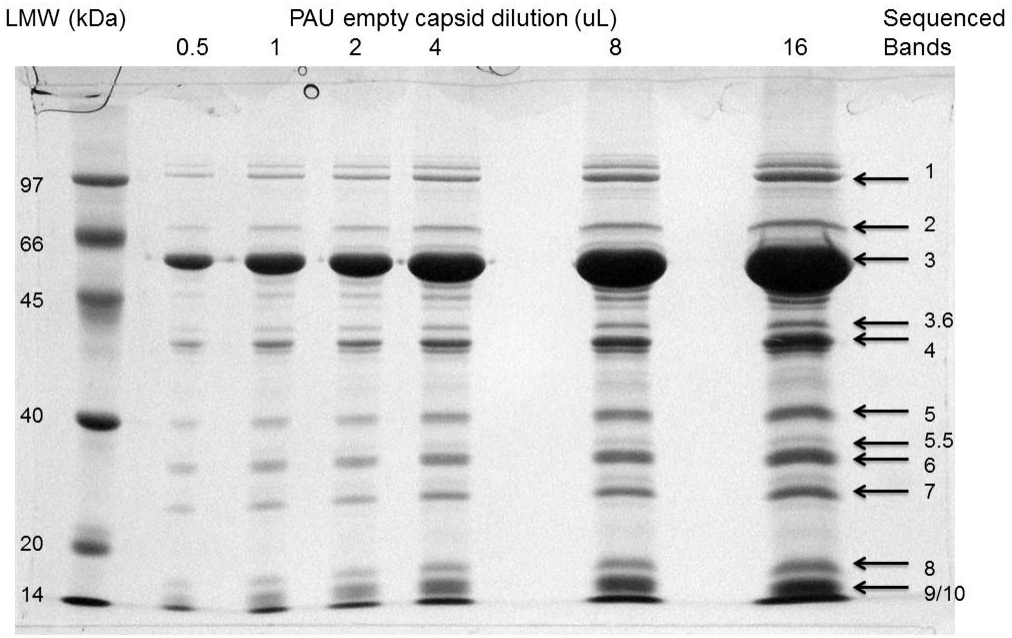
The MCP of PAU is gp156 (gene product of orf156), a 63.4 kDa protein with 585 amino acid residues (AAs). The N-terminal end is proteolytically processed, leaving a final 56.2 kDa protein with 521 AAs in the mature capsid. Removal of the first 64 AAs (the delta domain) turns the positively charged protein (isoelectric point (pI) 7.61) into a negatively charged final product

(pI 4.95) (Table 1), because the arginine-rich delta domain is highly positively charged (pI 9.76). While it is common to see that MCPs in tailed phages are negatively charged, such a transition of charge by proteolysis is uncommon. For example, the shift of pI caused by proteolysis in HK97 MCP gp5 (from 5.00 to 4.99) and in T4 MCP gp23 (from 5.31 to 5.05) is negligible. In the T=28 jumbo coliphage 121Q which I will introduce later, the pI slightly decreases from 5.74 to 5.25 after protease processing. The reason why the delta domain of PAU gp156 has very high pI is beyond the scope of this project. My guess is that the positively charged residues in the delta domain might regulate interactions with nearby delta domains from other gp156 subunits, and with proteins inside the capsid such as the protease, the scaffolding proteins and several internal proteins.

The PAU empty capsid contains at least 10 minor proteins in addition to the major capsid protein (Figure 5 & Table 1). For the other identified proteins, seven of these are in the 10–40 kDa size range and are present in ~100–300 copies/capsid. Three are in the 41–86 kDa range and present in ~60 copies/capsid. The seven smaller proteins are also proteolytically processed near their N-termini, at Glu-Ala sites that resemble the consensus sequence for T4 capsid processing (Keller, Kellenberger et al. 1985). Protein gp152 is present in both bands 5 and 6 with the same N-terminal sequence, probably a result of different processing by the protease because the 277 AAs protein has two Glu-Ala sites near the C-terminus started at G210 and G238. The 17.3 kDa protein gp142 is present in band 5 at about 30 kDa in addition to bands 9/10, which I have no explanation for except mishandling in sequencing. It is worth noting that the processing site in the MCP, Glu-Asn, is different from those of the other processed proteins. A possible explanation is that the two types of recognition sites allow the protease to selectively process substrates at separate stages during prohead assembly.

Figure 5 SDS PAGE of a series of PAU empty capsid dilution for Edman sequencing and band quantification

A series of dilutions of PAU empty capsid sample were electrophoresed. Those significant bands that were not tail protein bands recognized by comparing the phage protein with empty capsid protein lanes were identified by Edman degradation. We obtained reliable sequencing data from twelve bands. The bottom two bands were so close that the same result was acquired. The dilution series were used to provide more accurate estimate of band quantification.



SDS PAGE of PAU empty capsid protein dilutions

Table 1 Protein identification and gel quantification of PAU empty capsids

Band	ORF	Size/Size* (kDa)	N-terminal Cleavage	# of copies	pI/pI*	Notes	BLAST
1	9	85.5/-	-	41	6.89		
2	10	63.6/-	-	30	7.52		
3	156	56.2/63.4	LME|NYQ	1495	7.61/4.95		MCP
3.6	1	41.1/-	-	70	8.15		
4	270	38.5/40.1	LNE|ANE	176	7.83/8.30		
5	152 142	30.9/32.0 15.6/17.3	LYE|AIV LNE|ASG	97		Two bands with different intensity;	152: Scaffolding protein
6	66 152	24.6/30.6 30.8/32.0	KID|ALN LYE|AIV	199	7.94/9.19 7.92/7.86	Two bands with equal intensity	
7	199	19.6/20.5	LYE|AFV	164	7.89/9.30		
8	151	25.5/21.6	LFE|AVS	136	7.65/6.16		Protease
9/10	154 142	14.2/15.8 15.6/17.3	LNE|ASN LNE|ASG	546	7.06/4.91 7.54/6.97	Both proteins detected in 9 and 10;	

Notes:

1. The number of copies per capsid was calculated based on the fact that a T=25 capsid has 1495 copies of MCP.
2. Two proteins were detected in more than one band: gp152 in both band 6 and band 5, and gp142 in band 5 and bands 9/10.
3. The band quantification of band 5, 6 and 9/10 are the sum of quantification of intensity from the two proteins detected. The molecular weight of the proteins in band 5 was considered as 30.9 kDa. The molecular weight of the proteins in band 6 and bands 9/10 was considered the average of the two cleaved protein products detected.
4. The marker * indicates proteolytic processing.

2.3.2 Prediction of capsid protein function

Once the capsid proteins were identified, an immediate question to ask was their roles in the PAU capsid. This section covers what I learned from sequencing data, which leads to a conclusion that the minor proteins may fall into two categories. Later in this chapter I will present how we addressed this question more convincingly by a combination of biochemical manipulation and cryo-EM structure analysis.

Several years before I started the PAU project, sequencing of the PAU genome was finished in the Pittsburgh Bacteriophage Institute, founded by my advisor Dr. Roger Hendrix and another PI, Dr. Graham Hatfull, in the Department of Biological Sciences. The annotation of the PAU genome was hindered by lack of sequence similarity to other sequenced phage genomes. When the PAU genes were submitted for blast-p queries, only 39 out of all 292 predicted protein-encoding genes received at least one hit with the minimal E-Value of $1X10^{-10}$ (Table 2, Figure 6). In the 39 genes, three encode conserved phage proteins including the portal complex, the terminase large subunit and the head completion protein; the other genes encode either hypothetical proteins or proteins related to bacterial cell metabolism. Unfortunately, none of the 11 capsid proteins identified by Edman degradation is listed among the 39 genes with putative identities.

The genome shows a possible conserved cluster of head genes with consecutive genes encoding the portal protein, protease, scaffolding protein and MCP. Given that such a gene cluster exists in the PAU genome, genes encoding the protease and the scaffolding protein are expected to lie between the portal gene 147 and the MCP gene 156. HHpred (Soding 2005) was able to track down the sequence homology of gene 151 and identified it as the protease encoding gene. We then took advantage of the fact that the scaffolding protein is alpha-helix or coil rich to

search for the potential scaffolding protein. The high content of helices predicted by Phyre (Kelley and Sternberg 2009) in gp152 suggests that gene 152 encodes the scaffolding protein. The identification of gp151 and gp152 as processed proteins in the empty capsid further confirmed their identities. Therefore, the PAU genome still retains the classic cluster of head genes, but a few other genes are inserted in between.

The seven processed smaller proteins, including the protease and the scaffolding protein, may represent one type of capsid protein as internal proteins. The fact these proteins are proteolytically processed is strong evidence that they are inside the capsid, where access to phage protease activity is possible. Phage internal proteins are less well known compared with other capsid proteins. The T=13 prolate phage T4 has several small processed basic proteins, called IPs. During infection the IPs are injected into the host and bind the negatively charged T4 DNA as a protection against host restriction enzymes (Bachrach and Benchetrit 1974; Soding 2005; Thomas, Weintraub et al. 2012). Internal proteins were also found in the T=27 phage phiKZ (Thomas, Weintraub et al. 2012) and in the T=28 phage 121Q (Chapter 3 of this thesis) but their roles are not known. The high pIs of gp66, gp199 and gp270 suggest they may play a role similar to T4 IPs that bind and protect phage DNA during infection.

The three unprocessed proteins may represent another type of capsid protein, which decorate the exterior and interior of PAU capsid. The presence of the Glu-Ala protease recognition site near both N-terminus and C-terminus in gp9, but in neither gp1 or gp10, suggests that gp9 may be outside the capsid and gp1 and gp10 may be inside the capsid. The copy numbers per capsid of the three proteins are all 30 ~ 70 copies/capsid, calculated based on the gel quantification. This rough estimate suggests that each of the 60 asymmetric units may

have a single copy of all three proteins, which is in favor of the observation that decoration density of the PAU capsid is arranged around the pentamer.

The lack of detectable sequence homology and the presence of short unknown genes in the conserved gene cluster suggest that PAU has evolved and diverged quite far from phages that we are familiar with. While the seven genes that may encode the internal proteins appear scattered in the genome map, the three genes that encode the unprocessed proteins are nearby each other in the genome map, with seven short tRNA encoding ORFs between gene1 and gene9 (Figure 6). My speculation is that the genes that encode the three large unprocessed capsid proteins may be acquired all together by the phage recently, from a source that the sequencing community has not touched yet. Such an event may be related to capsid expansion and is worth further investigation.

Table 2 BLAST result of PAU open reading frames

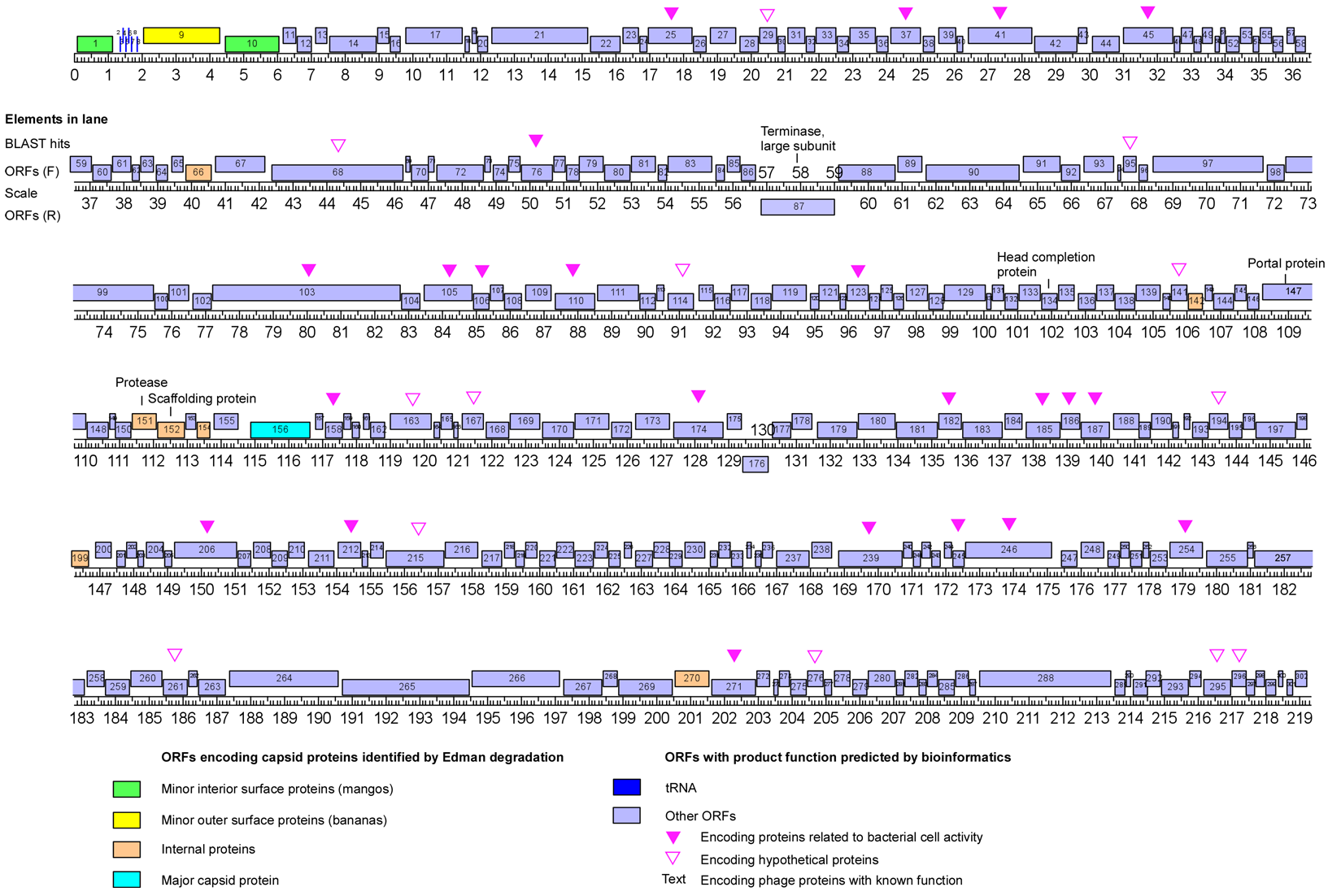
ORFs	BLAST hits	BLAST scores	Source of hits	ORFs	BLAST hits	BLAST scores	Source of hits
25	Thymidylate synthase	396	Bacterium	**167**	Hypothetical	210	Bacterium
29	Hypothetical	324	Both	**174**	Helicase	404	Phage
37	CMP deaminase	338	Bacterium	**182**	Thymidylate synthase	261	Bacterium
41	DNA topoisomerase 2-beta	660	Bacterium	**185**	dTDP-glucose 4,6-dehydratase	773	Bacterium
45	DNA topoisomerase 2-beta	425	Bacterium	**186**	dTDP-4-dehydrorhamnose 3,5-epimerase	356	Bacterium
68	Hypothetical	283	Bacterium	**187**	Glucose-1-phosphate thymidylyltransferase	981	Bacterium
76	Clamp loader subunit	413	Phage	**194**	Hypothetical	224	Both
87	Terminase large subunit	577	Phage	**206**	DNA ligase	466	Bacterium
95	Hypothetical	430	Both	**212**	Sugar isomerase	257	Bacterium
103	Bacterial surface protein	1271	Bacterium	**215**	Hypothetical	389	Phage
105	DNA polymerase	266	Phage	**239**	Helicase	416	Both
106	dCTP pyrophosphatase	216	Phage	**245**	Ribonucleotide reductase	299	Bacterium
110	DNA polymerase	205	Phage	**246**	Ribonucleotide-diphosphate reductase alpha subunit	1358	Bacterium
114	Hypothetical	235	Bacterium	**254**	Ribonucleoside-diphosphate reductase beta subunit	1253	Bacterium
123	Phosphoesterase	268	Both	**261**	Hypothetical protein	221	Bacterium
134	Head completion protein	215	Phage	**271**	dTDP-4-dehydrorhamnose reductase	229	Bacterium
141	Hypothetical	416	Both	**276**	Hypothetical	164	Phage
147	Portal protein	255	Phage	**295**	Hypothetical	173	Phage
158	Dihydrofolate reductase	258	Bacterium	**296**	Hypothetical	219	Bacterium
163	Hypothetical	459	Both				

Notes:

1. The minimal threshold to show a hit is at least one hit with the minimal E-Value of 1 X 10^{-10}.
2. The BLAST score is the largest score if more than one hit are received.
3. Source indicates what kind of organisms the BLAST hit belongs to.

Figure 6 Map of PAU genome

The map is generated by the program DNA Master by Dr. Jeffrey Lawrence and edited in Canvas. The identified capsid genes are indicated with different colors according to their locations in the capsid (see details in the later section in this chapter). Homology of predicted ORFs is labeled by text (phage genes), filled triangles (bacterial genes) and empty triangles (hypothetical). Predicted t-RNA encoding genes are indicated by thin blue bars.


PAU genome 219,373 bp
Elements in lane
BLAST hits
ORFs (F)
Scale
ORFs (R)
Terminase, large subunit
Head completion protein
Portal protein
Protease
Scaffolding protein
ORFs encoding capsid proteins identified by Edman degradation
Minor interior surface proteins (mangos)
Minor outer surface proteins (bananas)
Internal proteins
Major capsid protein
ORFs with product function predicted by bioinformatics
tRNA
Other ORFs
Encoding proteins related to bacterial cell activity
Encoding hypothetical proteins
Text
Encoding phage proteins with known function

2.4 STRUCTURE OF THE PAU CAPSID

2.4.1 Predicted HK97 fold in the PAU major capsid protein

PAU is probably an isolated individual from other known phages. It is the only T=25 phage, the only *Sphingomonas paucimobilis* phage and the only phage originally isolated from silkworm extract. Only a few of the phage genes have homologues found by BLAST, but the major capsid protein gp156 is not one of them. The first problem for me to address about PAU capsid is whether the major capsid protein of PAU, despite poor sequence similarity, adopts the HK97 fold that is shared by all tailed phage MCP structures that have been studied.

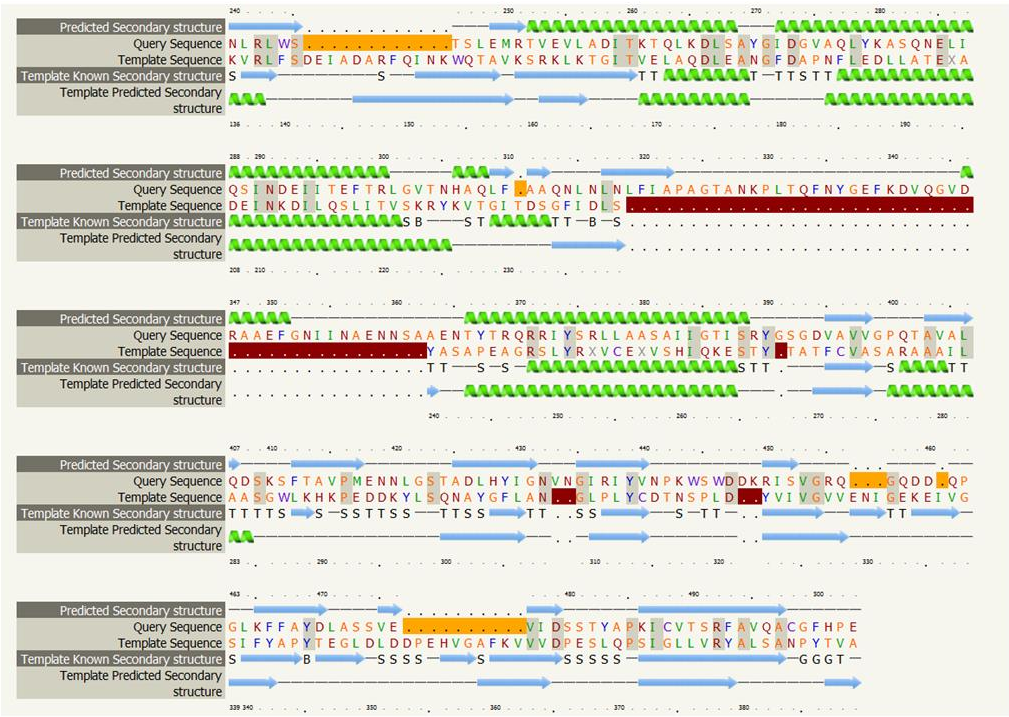
The secondary and 3D structure of gp156* (*: proteolytically processed) predicted by Phyre2 webserver (Kelley and Sternberg 2009) suggests the gp156 has the HK97 fold. Prediction suggests that the secondary structure arrangement of the C-terminal, including residues 310-382 and 427-567 in gp156, matches the residues 155-389 of T4 vertex protein gp24 with high confidence (Figure 7A, 1st row). For the matched region, counterparts in gp24 can be found for most predicted α-helices and β-strands in gp156. The matched region of gp24 employs the core HK97 fold that contains an axial domain (A domain) with several short α-helices and β-strands, and a peripheral domain (P domain) that consists of a long helix and a few long strands (Fokine, Leiman et al. 2005). A homology 3D model of PAU gp156* predicted by Phyre2 shows the two matched regions adopt the HK97 fold, forming the A domain and P domain that account for the majority of the continuous layer of capsid shell (Figure 7B & C). Structure prediction of the gp156* by another web server, I_TASSER (Yang, Yan et al. 2015), provided a very similar model which was also based on homology to T4 gp24. Therefore, structure prediction suggests that PAU MCP gp156 is likely to be another member of the conserved HK97 MCP family.

Three parts of PAU gp156* do not match with the gp24 structure with high confidence, including the N-terminal 239 residues, residues 383-426 between the two matched regions and a short 16 residue C-terminus. The most significant part is the residues at the N-terminal, which may form the protruding knob near the edge of hexamers. It may also interact with adjacent subunits, playing a role similar to that of the N-terminal elongated loop in HK97 gp5 and the insertion domain in T4 gp24.

Figure 7 Structure prediction of PAU major capsid protein gp156.

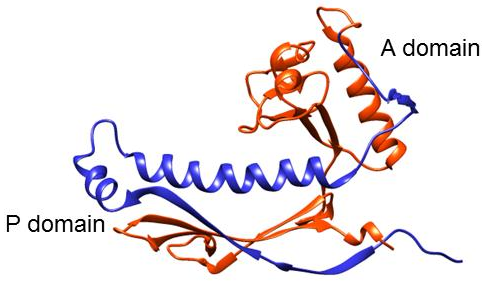
A. Secondary structure prediction of proteolytically processed gp156 by Phyre2. Query sequence is the gp156* sequence so the AA numbers are off by 64, the length of the delta-domain. Template sequence is the T4 gp24 sequence. From top to bottom, the three lanes of secondary structure are predicted secondary structure of gp156, secondary structure of gp24 (PDB ID: 1yue) and crude predicted secondary structure of gp24 replaced with matched gp156 sequence. Green helices indicate α-helices and blue arrows indicate β-strands. The secondary structure arrangement of residues 310-382 and 427-567 in gp156 match the residues 155-389 of T4 vertex protein gp24 with high confidence.
B. Crystal structure of the matched regions in gp24. The region forms the A domain and P domain that are the signature of the HK97 fold. Blue: the region of gp24 matches residues 310-382 in gp156*; red: the region matches residues 427-567 in gp156.
C. Predicted 3D structure of the matched regions in gp156. Gray: unmatched residues 383-426.

A



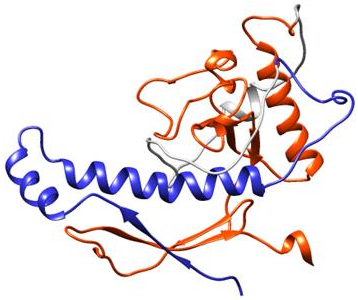
Phyre2 prediction of PAU gp156*

B



T4 gp24*, residues 155-389

C



PAU gp156*, residues 310-567

2.4.2 EM Structure of the PAU empty capsid

Through collaboration with Dr. James Conway and Dr. Katerina Toropova in Department of Structural Biology, University of Pittsburgh, we obtained a 9 Å density map of the PAU empty capsid. The surface features shared between the cryo-EM model of the empty capsid and the model of the PAU phage suggests that the empty capsid we isolated is expanded and mature (Figure 8A & C). An expanded mature capsid could either be a by-product of phage assembly that has never had any DNA packaged, or a once DNA-filled head that has somehow lost its DNA during purification, likely as a result of damage to the capsid. The isolated empty capsid is more likely to be the first species, because all the particles appear intact under the electron microscope. The cryo-EM reconstruction of broken particles would be unlikely to achieve a 9Å resolution because of the heterogeneity of the particles.

The PAU capsid has a T-number of 25 and a rugged outer surface with a vertex-to-vertex diameter of 1300Å (Figure 8A). The icosahedral capsid is built by 240 hexamers and 11 pentamers (the portal complex replaced one pentamer at a vertex) using 1495 MCP subunits. Each subunit has a protruding knob near the edge of the hexamer or pentamer. Each of the 60 hexamers that is adjacent to one of the 12 pentamers has arch-like density sitting on top, spanning from the center of the hexamer to one of its knobs (Figure 8A & E). A narrow hole is observed at the center of the other hexamers that do not have the arch-like density. The outer surface feature of the pentamer is similar to that of those hexamers without the arch-like density. An additional structure beneath the pentamer extends about 9 nm, forming a pentameric structure on the interior, which in the DNA-containing particles displaces DNA layers near the pentamer (DNA data not shown) (Figure 8B & F). Both the T=25 triangulation number and the distinctive arrangement of minor capsid proteins around the pentamer are unique among characterized

phage capsids (Figure 8D). Note that each of the 60 arch-like densities on the capsid must be in the same orientation with respect to the adjacent pentamer. If that were not the case their density would be blurred out during the 60 fold averaging. The arch-like density and the pentameric density were named “banana” and “mango” in our lab, because of the shape and color of their structures when I first presented the data. I use these names throughout this chapter because the names are convenient and vivid to describe the extra density.

Figure 8 cryo-EM structure of PAU empty capsid

A. Surface view of the empty capsid at 9 Å. The coloring of the density is determined by the radii to the center of the icosahedral capsid. A 200% magnification of the circled area is available in Figure 8E, which covers the banana (the arch-like density) on hexamers surrounding the pentamer.

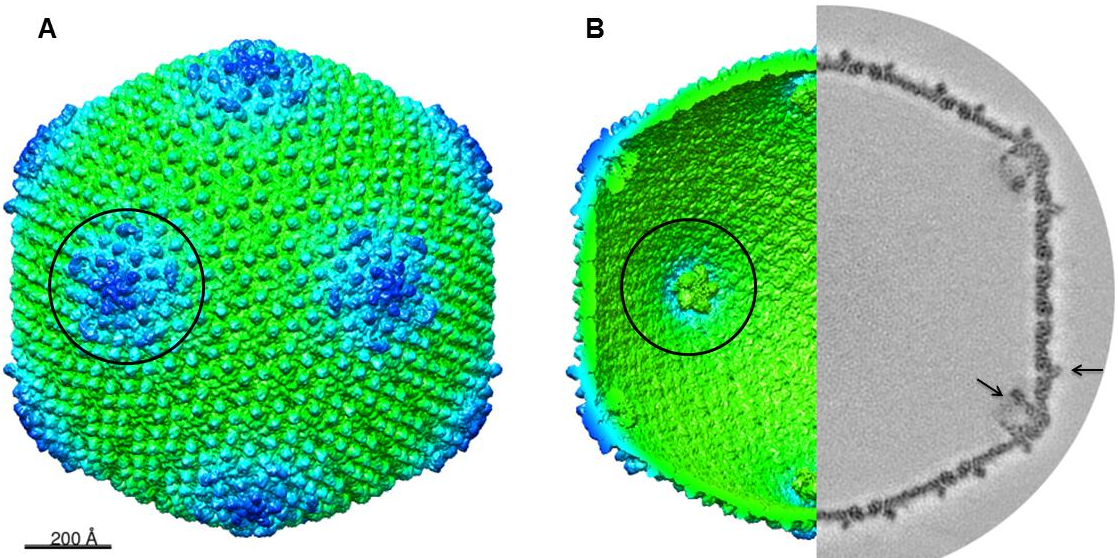
B. Interior surface view of the empty capsid and a slice of the density map. A 200% magnification of the circled area is available in Figure 8E, covering the mango (the internal pentameric density) on the interior surface of the pentamer. The slice of the density map represents the actually density at the plane, while the surface view of the density map uses a cutoff to show the volume with density equal or higher. Arrows point to the position of the banana and mango density on the slice.

C. Surface view of PAU phage at 20 Å.

D. Schematics of a triangular facet of the PAU capsid. Orange arc: banana density; blue dot: mango density; shaded area: an asymmetric unit of the T=25 capsid; numbers indicate the number of steps between pentamers: h values in the nomenclature of Caspar and Klug (Caspar and Klug 1962), used to calculate the T-number; small shaded penton at the vertex of the triangle: positions of the penton vertex of the PAU capsid.

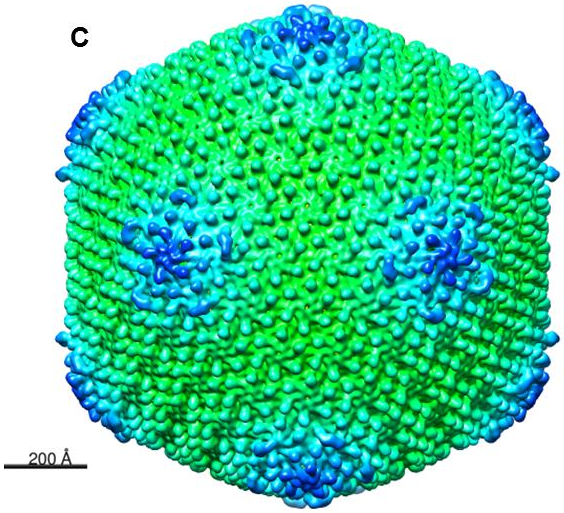
E. A 200% magnification of the circled area in Figure 8A. The arrow points to a banana density.

F. A 200% magnification of the circled area in Figure 8B.

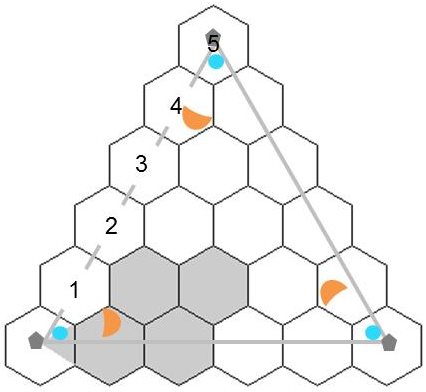


Surface view of the empty capsid

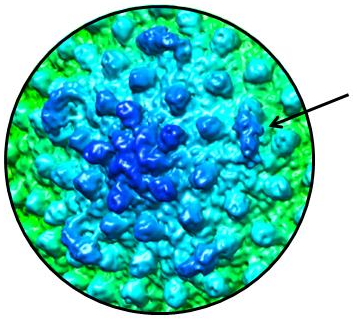
interior and slice of the empty capsid



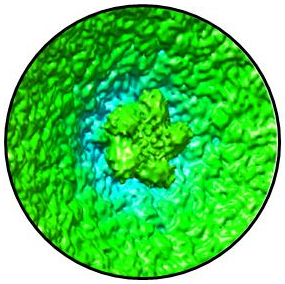
Surface view of PAU phage



Schematic of the PAU triangle



Outer surface near the penton



Interior surface near the penton

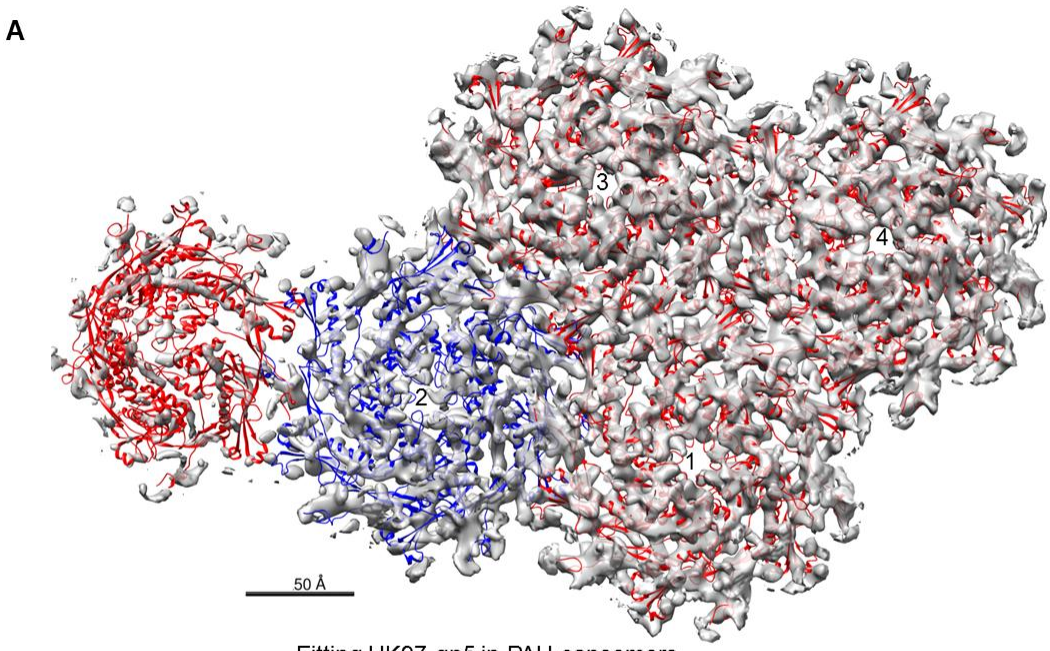
2.4.3 The capsid shell formed by the HK97-fold

Stronger evidence that gp156 adopts the HK97 fold was obtained from fitting the X-ray coordinates of mature HK97 head (PDB ID: 1OHG) (Figure 9B) into the cryo-EM density map of PAU empty capsid using UCSF Chimera (Pettersen, Goddard et al. 2004) (Figure 9). A density map at a resolution higher than 10Å is able to show the trace of α-helices so the distinctive 40 Å long helix in HK97 MCP gp5 is a useful indicator of the quality of the fitting. Five identical gp5 chains of the HK97 pentamer (1OHG, chain F) were individually fitted into the pentamer density of the PAU empty capsid density map. Additionally, the HK97 hexamer (1OHG, chains A-G) was split into 6 chains and then the chains were individually fitted into the hexamer density near the 2-fold symmetry axis (hexamer #1 in Figure 8A) in the map. The rearranged hexamer structure after fitting was also fitted into the other three hexamers in the same asymmetric unit.

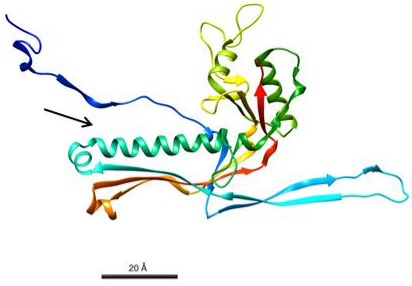
The gp5 structure fits well with the central hexamers that do not have direct contact with a pentamer (hexamers #1, #3 and #4 with red ribbon structure fitted in Figure 9A) (Figure 9C). All three hexamers have similar local symmetry and the outer surface of their density is relatively flat. A closer view of the fitted a gp5 subunit in a middle hexamer clearly shows that there is a long rod-like density that precisely accommodates the long helix in gp5 (Figure 9D). There is also density that corresponds with the A domain and the long β-strands in the P domain. Therefore, it is safe to conclude that the PAU major capsid protein gp156 adopts the HK97 fold despite considerable sequence divergence.

Figure 9 Fitting the crystal structure of HK97 mature head structure (1OGH) into the PAU empty capsid.

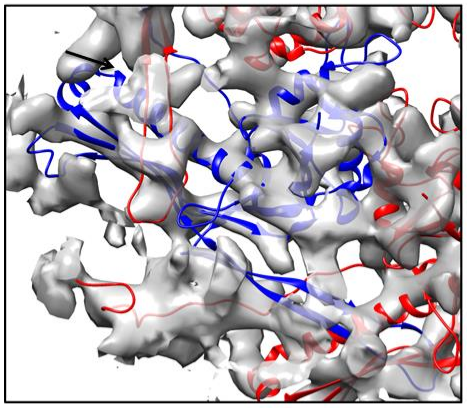
A. Fitting HK97 MCP gp5 into the PAU empty capsid density. The density includes one pentamer and four different hexamers. The four numbered hexamers and one subunit in the pentamer form the asymmetric unit of the T=25 capsid. The coordinates of the HK97 pentamer and hexamer were split into individual chains and fitted into the pentamer and hexamer #1 with UCSF Chimera. The resulting hexamer coordinate was then fitted in the density of hexamers #2, #3 & #4 as a whole. The coordinates fitted in hexamer #1 are colored blue. The cyro-EM density was zoned 7 Å around the coordinates, resulting a clearer visualization of the fitted density but losing track of the extra density that cannot be fitted with the gp5 structure.
B. Structure of HK97 gp5. The arrow points to the long helix.
C. A closer view of a single fitted gp5 subunit colored in blue in hexamer #1. The arrow points to the long rod-like density fitted by the long helix in the gp5 P domain. The image is a 2 X magnification of the squared region in Figure 9D.
D. Fitting gp5 in hexamer #1. Using hexamer #3 or #4 instead of #1 resulted in similar fitting outcome.
E. Fitting gp5 in the pentamer.



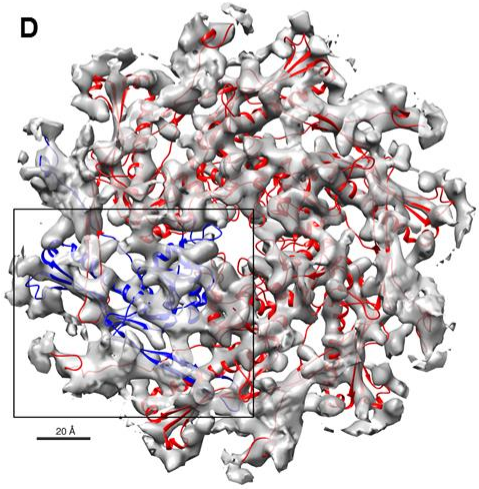
Fitting HK97 gp5 in PAU capsomers



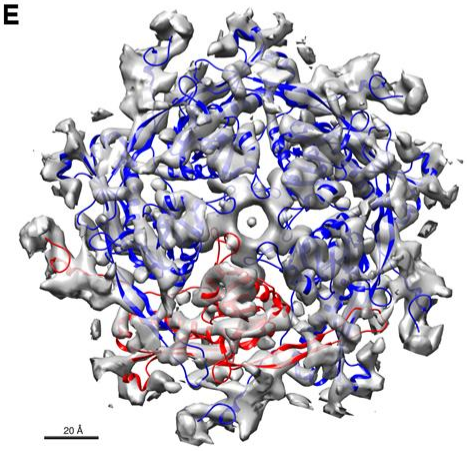
HK97 gp5 structure



Close view of a subunit in hexamer #1



Fitting gp5 in hexamer #1



Fitting gp5 in the pentamer

The fitting of the pentamer and of the hexamer adjacent to the pentamer (hexamer #2) is less satisfying. The density of the pentamer and of hexamer #2 is weaker and more blurry compared to the density in the other hexamers at the same contour level, suggesting that the structure around the pentamer is more flexible. Despite the relative low fitting quality, the five subunits still formed a symmetric pentamer after individually being fitted into the density and a long rod-like density is very close to the long helix of each subunit (Figure 9E).

The same hexamer structure that fits well with hexamers #1, #3 and #4 apparently did not perform well with the density of hexamer #2 as a solid boy fit. Not all of the six long helices precisely match the nearby long rod-like density. The greatest discrepancy between the hexamer coordinates and the density map is displayed by the fitting of chain B which is located at the subunit under the banana. Most of the long helix in chain B is completely off the rod-like density, except the end of the helix near the A-domain (Figure 10A). Note that an extra piece of density from the knob or the banana covers the other end of the helix in Figure 10A, making it difficult to observe the distance between the helix and the rod-like density in the top view. The apparent mismatch between the density of hexamer #2 and the coordinates that fit the other three hexamers suggests that the structure of hexamer #2 may be noticeably different from the others.

The conformational variation between the four hexamers in the T=25 geometry is demonstrated by superimposing the fitted gp5 coordinates and the density from different hexamers (Figure 10C & D). Superimposition of the hexameric coordinates fitted with hexamers #1, #3 and #4, or the density of these hexamers, show no noticeable difference (data not shown). The fitting of six copies of individual gp5 subunits in the hexamer #2 density generated a new hexameric structure with better local fitting of the long helices to the rod-like density. However, the new hexameric coordinates is apparently asymmetric just by comparing the positioning of the

loop of each subunit near the center of the hexamer (Figure 10B). Superimposing the asymmetric structure and the hexamer coordinates obtained from fitting gp5 in hexamer #1 shows that two pairs of chains have noticeable difference (Figure 10C). The long helices of chains A and B (chains named using the naming convention for 1OHG) from the two superimposed coordinates could not be matched as closely as the other four chains did. Superimposing the density of hexamer #1 and the density of hexamer #2 also shows that the long rod-like density of subunit B displayed a significant mismatch of about 10Å (Figure 10D). However, the rod-like density of subunit A is weak and indistinguishable from the nearby density at any contour level. The position of the rod-like density of other subunits is very close for each pair.

An important principle of the quasi-equivalent construction of an icosahedral virus capsid is that chemically identical protein subunits adopt slightly different conformations that are determined by their specific environments (Caspar and Klug 1962). For the 24 hexamer subunits in the T=25 PAU capsid asymmetric unit, our 9Å capsid structure shows no significant difference between the density of the hexamer subunits except two subunits in hexamer #2. The conformation varieties of the two subunits in hexamer #2 are likely caused by the molecules they interact with. While the 22 similar subunits mainly interact with other hexamer subunits, subunit A in hexamer #2 is the main subunit that interacts with the pentamer and subunit B in hexamer #2 is the main subunit that interacts with the banana density.

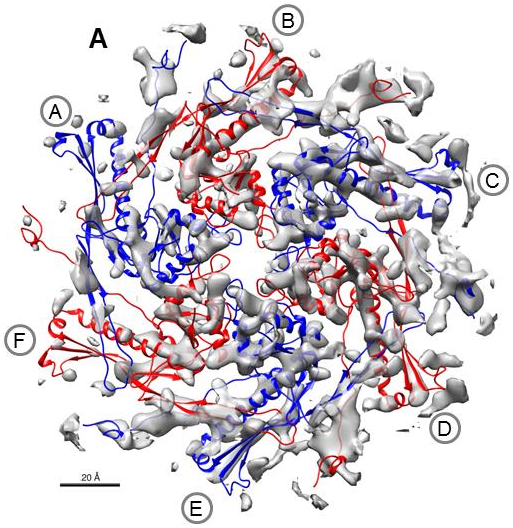
Figure 10 Conformational variations in different hexamers

A. Fitting the hexamer coordinates of hexamer #1 to hexamer #2 that has the banana density. Circled capital letters indicate the different chains named the same as the hexamer chains in 1OHG. Chains A and F interact with the pentamer. The banana mainly interacts with chain B (Figure 10C). The long helix of chain B is largely off the rod-like density. Most of the banana density is masked so it will not cover the hexamer density, but still a small piece of it remains where the long helix of chain B and the E-loop of chain C cross each other.

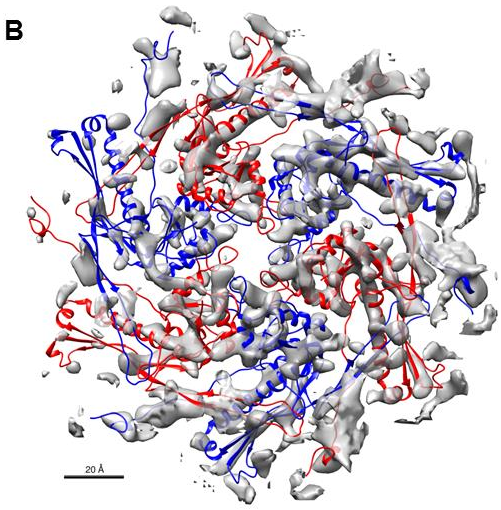
B. Fitting individual gp5 subunits to hexamer #2. The fitting is better for each subunit but the hexamer coordinates are no longer symmetric.

C. Superimposition of the two hexamer coordinates in Figure 10A (colored in cyan) and Figure 10B (colored in purple). The yellow transparent ellipse in the left picture marks the location and shape of the banana density. The right figure is an interior view of the same superimposition that is rotated 135˚ counter-clock wise along the x-axis. Chains A and B differ the most, based on the position of the long helices.

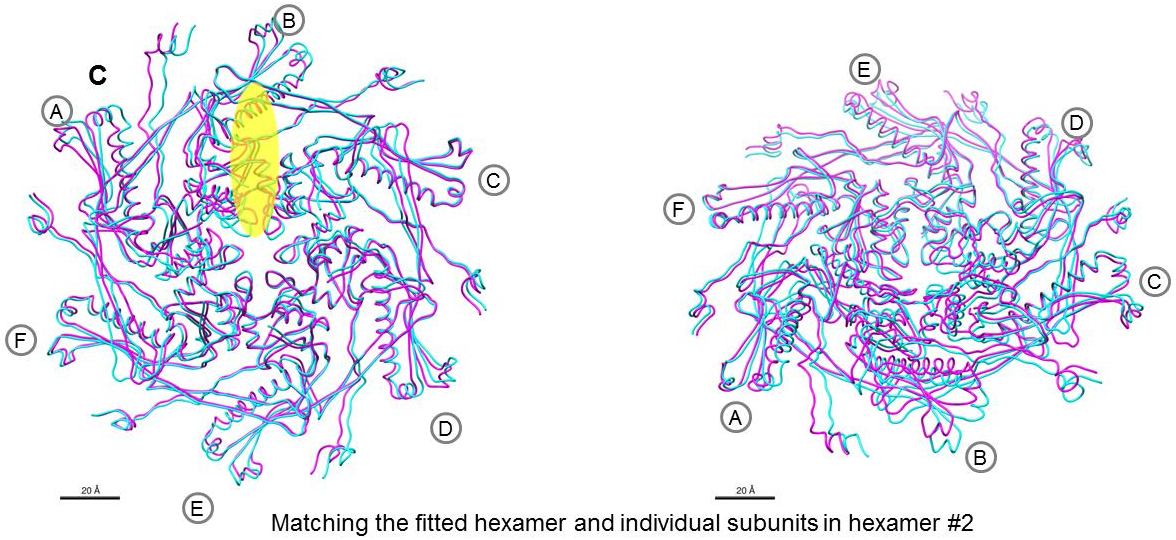
D. Fitting the density of hexamer #2 (colored in purple) to hexamer #1 (colored in cyan with 70% alpha). The long helices of the hexamer #1 coordinates are shown in red to help locating the long-helices. The right picture is an interior view of the same density rotated 135˚ counter-clock wise along the x-axis. The density of A and E are much weaker in hexamer #2 than the density at the equivalent areas of hexamer #1. The position of helix of B shows noticeable difference between the two hexamers.



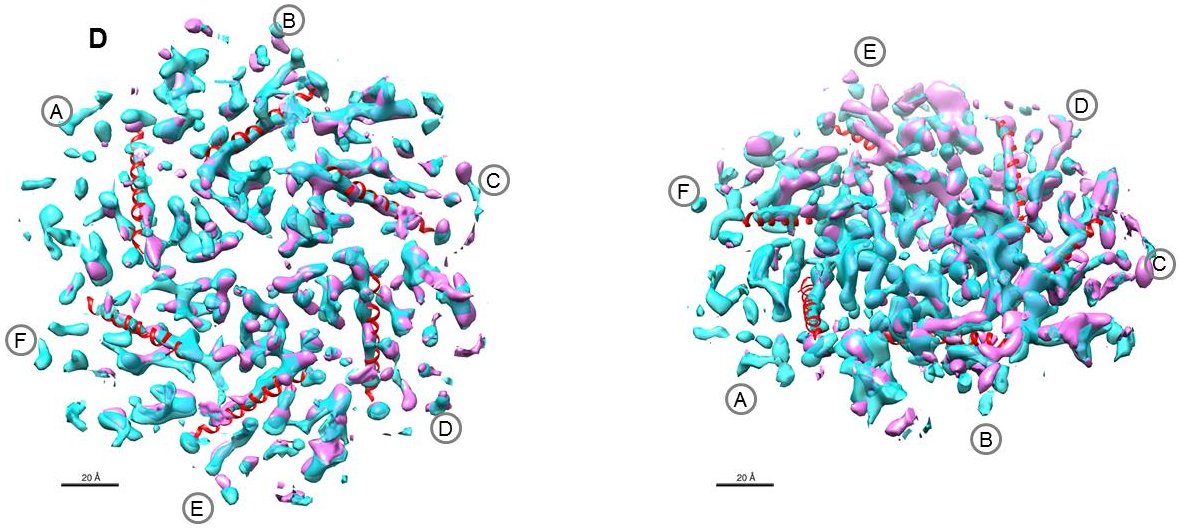
Fitting a hexamer structure in hexamer #2



Fitting individual gp5 in the hexamer #2



Matching the fitted hexamer and individual subunits in hexamer #2



Fitting density of hexamer #2 to hexamer #1

2.4.4 Additional density beyond the HK97 fold

The observation that the gp5 crystal structure fits nicely into the PAU capsid density map also made it possible to identify the extra features of the PAU density map that did not fit with the gp5 coordinates. The density within 7Å of the fitted gp5 coordinates is discussed in the previous section, and the other density which is more than 7Å away from the gp5 coordinates is discussed in this section.

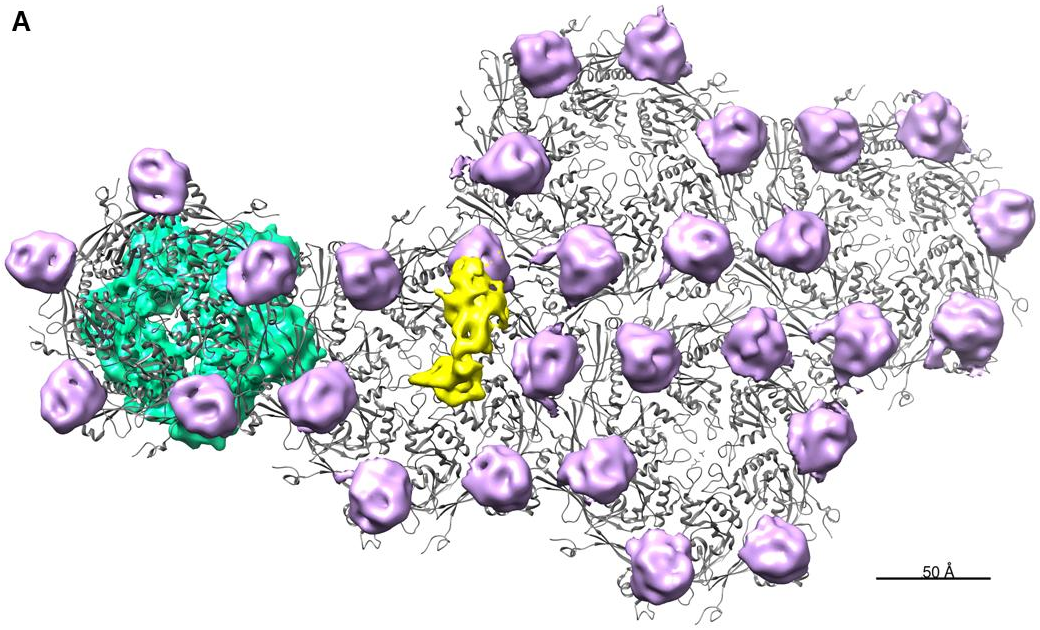
Removal of the density near gp5 left three groups of extra density on the PAU capsid density map (Figure 11). On the outer surface, there is the banana density sitting on mainly one MCP subunit of the hexamer near the pentamer, and the protruding knob on each subunit. The only extra density on the interior surface of the capsid is the mango density. The more blurry appearance of the banana density compared to the mango or the knob at the same contour level suggests that the banana density is more flexible.

It is evident that this knob density is formed by the N-terminus of gp156*. On one hand, no potential surface protein with copies/capsid ratio close to gp156 has been identified. On the other hand, the processed PAU gp156 (521 residues) is much larger than the processed gp5 (285 residues). The N-terminal region (the first 239 residues in the processed form or residues 65-303 before proteolysis) is not included in regions that were predicted to form the HK97 fold. Because the density sits on the contacting area of the E-loop and P-domain of two nearby subunits in the same capsomer, it may play a role in stabilizing the capsomer.

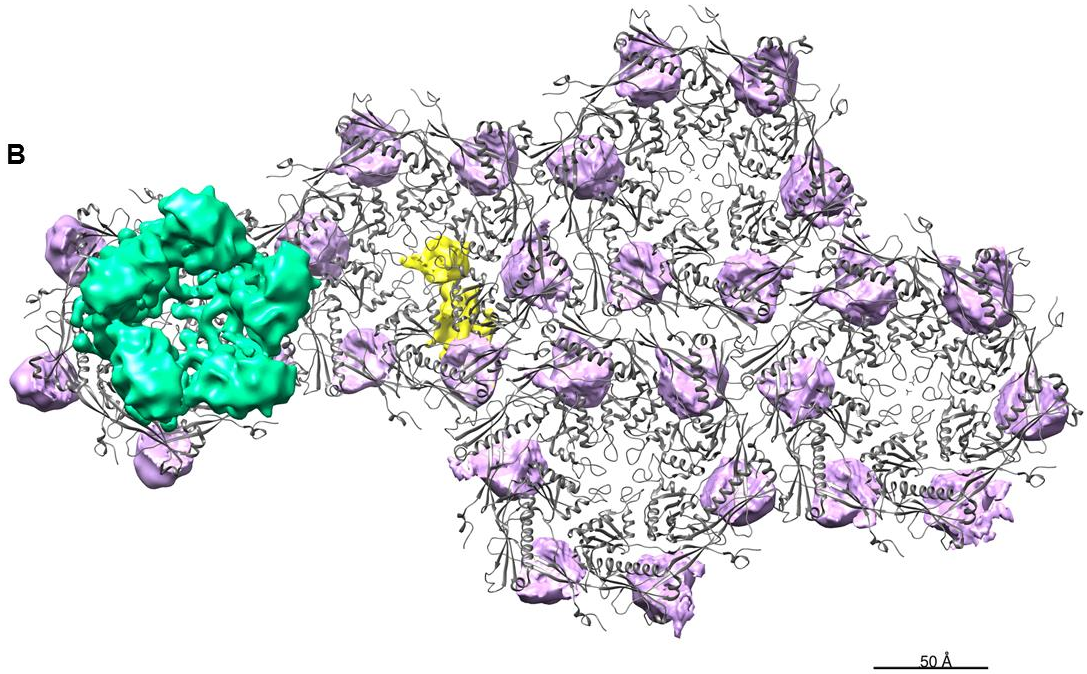
Figure 11 Density that is not fitted by the gp5 coordinates

A. Surface view of the density that is not fitted by HK97 gp5. Only the density of one pentamer and four hexamers are considered. The density that is within 7Å of the gp5 coordinates, which is shown in Figure 9 and Figure 10, is masked. The remaining density falls into three categories: the banana density (yellow), the mango density (green) and the protruding knobs (purple). The contour levels of the three different density groups are set to the same. The fitted gp5 coordinates are in gray.

B. Interior surface view of the density that is not fitted by gp5. The model is rotated about 180˚ along the x-axis compared to the model in A.



Density not fitted by HK97 gp5 coordinates



The Mango density on the interior surface

2.5 ORGANIZATION OF THE MINOR CAPSID PROTEINS

2.5.1 Overview of the chemical manipulation of PAU capsids

The banana and mango density on the PAU capsid form a unique arrangement of decoration density which is not observed in other phages. An important step to study these extra densities is identifying structural proteins that build the banana and mango density. Because genetic techniques for manipulating PAU and *Sphingomonas paucimobilis* are not currently available, biochemical manipulation of the PAU capsid became my only viable option. A decoration protein can be identified if a proper condition can extract it from the PAU capsid while leaving the capsid shell intact for EM reconstruction.

I treated both PAU empty capsids and phages with urea, guanidine chloride and heating. Urea treatment of the empty capsid successfully removed internal proteins and the banana protein at different concentrations of urea, but left the capsid shell and the mango density intact at the highest urea concentration that I could prepare in room temperature. Heating the empty capsid over 70˚C reduced the amount of banana and mango density on the capsid but not the internal protein, and the empty capsid remained intact. Guanidine chloride at low concentration could remove internal proteins from the PAU empty capsid, but the capsid was unstable at higher concentration. The PAU phage quickly released DNA with any of the three treatments, suggesting that high pressure from the densely packed DNA inside makes the capsid unstable in these conditions (data not shown).

2.5.2 Urea extraction of minor capsid proteins from the capsid

Dissociating minor capsid proteins from the PAU capsid requires a moderate denaturing environment that interferes with the weak interaction between the capsid shell and the minor capsid protein, but does not affect the strong interaction between the MCP subunits. We used a series of concentrations of urea, a widely used mild protein denaturing reagent for relaxing the hydrophobic interaction in protein (Bennion and Daggett 2003), to extract minor capsid proteins from the empty capsid. The PAU virion broke and released DNA at 2.5 M urea or higher, and so was not suitable for the experiment. Fortunately, the empty capsid was stable even in saturated urea at room temperature. The purified PAU empty capsid was treated with a series of concentrations of urea in TM buffer at room temperature for 1 hour. The sample was pelleted by ultracentrifugation, separating the dissociated minor capsid proteins in the supernatant and the capsid in the pellet. The resuspended pellet and the TCA-precipitated supernatant were then subjected to SDS-PAGE and/or TAMg agarose gel electrophoresis.

The SDS gel of the urea-treated PAU capsid shows that the previously identified minor capsid proteins respond differently to urea treatment (Figure 12A). When the empty capsid was treated with 3M urea or more, the seven small and proteolytically processed proteins no longer associated with the pelleted empty capsid, suggesting that 3M urea was enough to remove these proteins from the capsid. The largest protein, gp9, started to leave the capsid in presence of 5 M urea. The other two proteins gp1 and gp10 remained on the capsid shell even in saturated urea (data not shown). Because urea enhances the aqueous solubility of peptides and weakens the inter- and intra-molecular hydrophobic interaction in a protein complex, hydrophobic interaction may be the main force that binds the small proteins, gp9, and the capsid shell together. Other forms of inter-molecular interaction, such as salt-bridges, may strengthen the interaction between

the shell and gp1/gp10. The stability of gp1 and gp10 at high concentrations of urea suggests that these two proteins may be essential to the integrity of the capsid shell.

The agarose gel of the urea-treated capsid suggests that 5 M or higher concentration of urea changes the outer surface of the PAU capsid (Figure 12B). In TAMg agarose gel electrophoresis, the mobility of a phage capsid depends on both the size and shape of the capsid, and the outer surface charge of the capsid. All urea-treated capsids formed recognizable bands in the TAMg agarose gel, suggesting that there was an intact and uniform capsid species in each sample. The negative stain EM images of urea-treated capsids also suggest that the capsid shell was not damaged by the high concentration of urea, because the urea-treated capsid displays no apparent structural difference compared with the control capsid (Figure 12C).

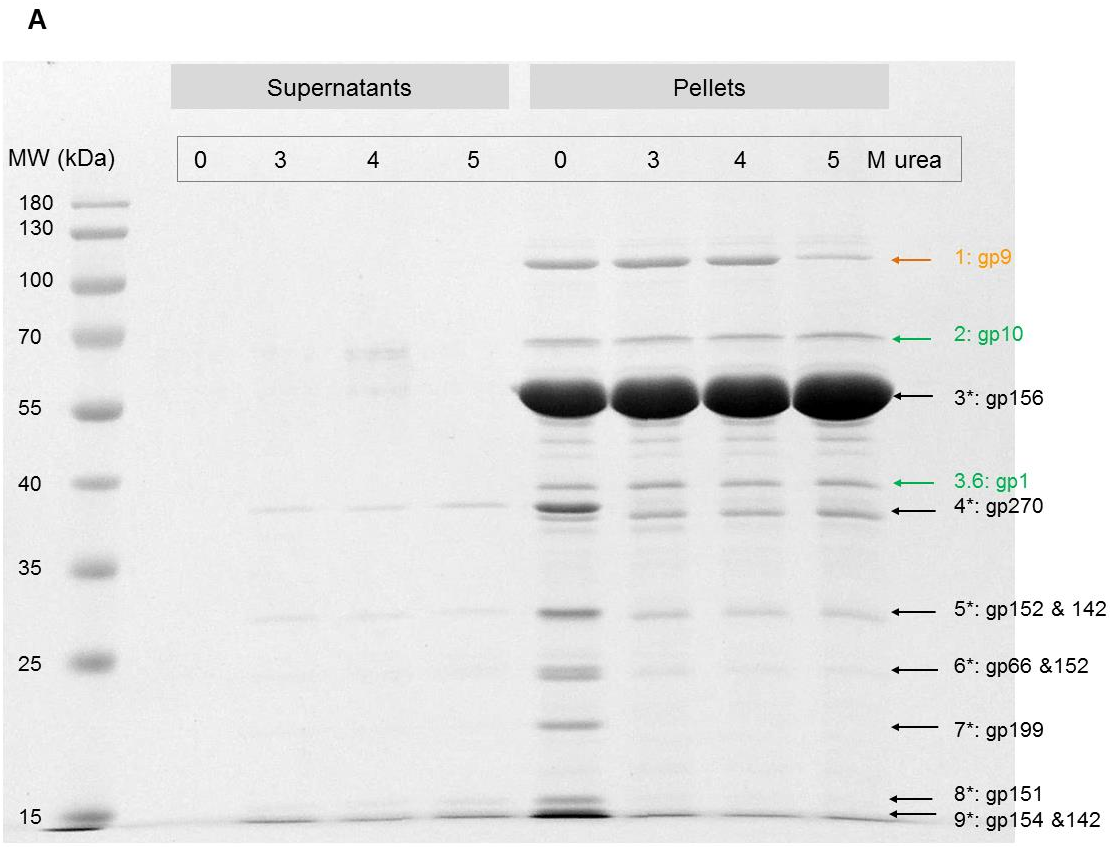
Because the capsid maintains its integrity after urea extraction and the size and shape of the capsid are (almost) unchanged, the only thing that can cause a dramatic shift of particle mobility in the agarose is the change in surface charge. The agarose gel shows that the mobility of the capsid treated with 4 M urea is the same as the mobility of the control capsid, but with 5 M or higher concentration of urea, the mobility of the capsid becomes higher (Figure 12B). This observation argues that the small proteins that were dissociated with 3M urea are inside the capsid, and gp9 which began to leave the capsid with 5 M urea is a banana protein on the outer surface of capsid. The gp9 surface exposed to the solvent is likely to be positively charged, so the net surface charge of the capsid turns more negative upon losing gp9, leading to higher mobility in electrophoresis in the range of 5 – 8 M. Capsids treated with higher concentration of urea ran faster on the agarose gel, suggesting that the dissociation of gp9 was not complete in 5 – 8 M urea.

Figure 12 PAU empty capsids treated with urea

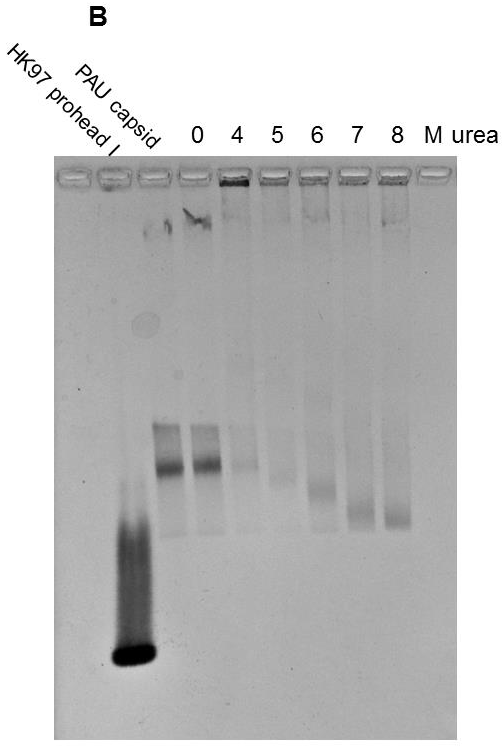
A. SDS gel of the supernatant and the pellet of centrifuged PAU empty capsid samples treated with different concentrations of urea. The urea concentration was indicated on top. The protein in the supernatant was precipitated with 10% TCA. Equal percentages of the total supernatant protein and the total pellet protein were loaded into the gel. The supernatant protein lane did not accurately reflect the protein lost in the pellet protein sample. The likely reasons could be that the recovery rate of TCA precipitation was poor and some protein bound to centrifuge tubes/eppendorf tubes. Bands that were identified by N-terminal sequencing are labeled at the right side in black: MCP and the small proteins, yellow: gp9, and green: gp1 and gp10. Two bands remain in the capsid after the urea treatment, one slight below band 4 and the other slight below band 5. The proteins in these two remaining bands may not be capsid protein.

B. TAMg agarose gel of the resuspended pellet. The urea concentration is labeled on the top of the gel. The mobility of phage capsids in the agarose gel depends on the shape and the surface charge of the capsid. An Hk97 prohead I sample and a PAU empty capsid sample were loaded as control. Some protein was trapped in the wells, which could be aggregated capsids. The major band shifts to the bottom as the concentration of urea increase. The formation of a clear band suggests that the capsid structure is uniform, i.e., undamaged after the urea treatment.

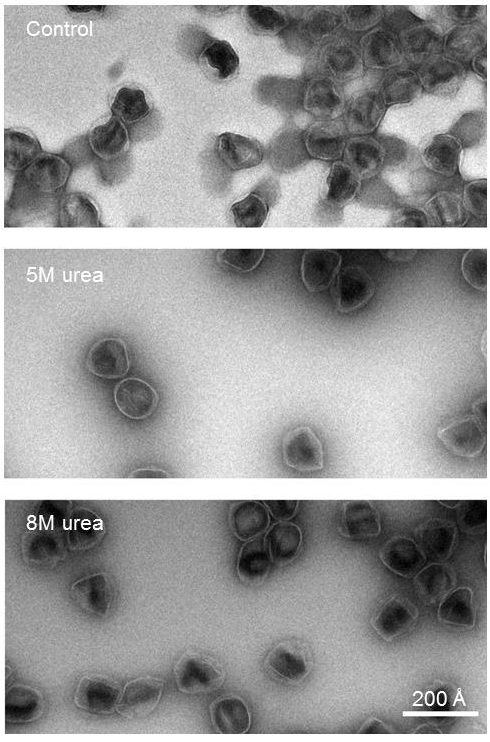
C. Negative stain images of urea-treated PAU empty capsids. The PAU empty capsids were treated with 0M, 5M, and 8M urea.



SDS gel of the supernatant and the pellet of centrifuged urea-treated PAU capsids



Agarose gel of urea-treated capsids



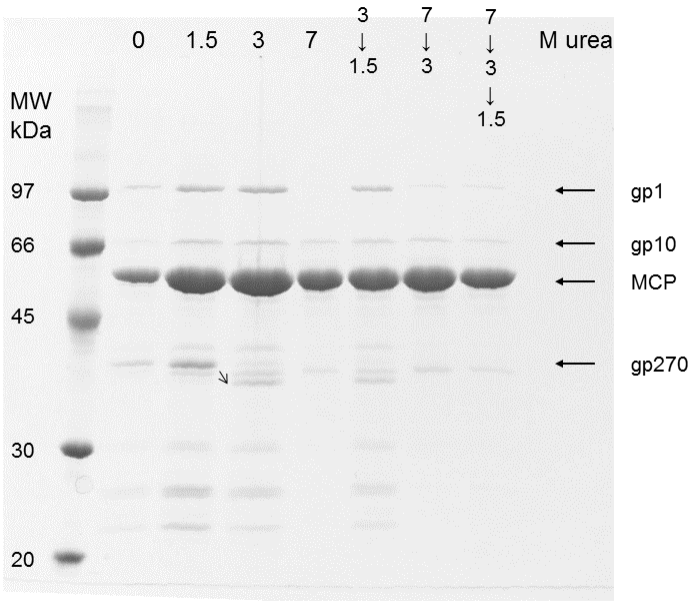
Negative stain EM images of urea-treated capsids

One explanation for the incomplete dissociation of gp9 in 5~7 M urea is that there is a chemical equilibrium during the dissociation process. Another less likely explanation is that the 60 copies of banana density are not identical and some density leaves earlier than the other. We tested the first hypothesis by checking if the urea extraction of the minor capsid protein is reversible. The PAU empty capsid was initially treated with 3 M or 7 M urea on bench for 2 hours to dissociate the minor capsid proteins. The reaction was diluted by TM buffer to reduce the urea concentration and then incubated on the bench overnight. If the dissociation of a minor capsid protein is reversible, the incubation at lower urea concentration will allow the reassociation of minor capsid proteins back to the capsid. At each step, an equal amount of PAU capsids was taken out, concentrated by ultracentrifugation, and prepared for SDS gel electrophoresis.

The lane of the capsids treated with 7 M urea contains no gp9 nor the small protein bands (except the two unidentified bands below bands 4 and 5), suggesting that the capsids treated with 7 M urea completely lost gp9 and the small proteins (Figure 13). The lane of the capsids treated with 7 M urea and followed by a dilution to 3 M urea contains a faint band of gp9, but no bands of the other lost proteins. The lane of the capsids incubated with a further dilution to 1.5 M urea shows no change in the band pattern. The reappearance of the gp9 band suggests that gp9 partially reassociated to the capsid when the urea concentration was lowered to 3 M, so the dissociation of gp9 by urea is at least partially reversible in the condition that we tested. The capsids did not regain any lost small protein when the reaction was incubated in 1.5 M urea in which the small proteins stay in the capsid, so the dissociation of the small proteins may not be reversible.

Figure 13 SDS gel of PAU capsids treated with decreasing concentrations of urea

The lanes are labeled by the urea concentration used for the sample loaded. The PAU empty capsids were initially treated with 0 M, 1.5 M, 3 M or 7 M urea for 2 hours. The 3 M urea reaction was diluted to 1.5 M. The 7 M urea reaction was diluted twice to 3 M and 1.5 M. Each dilution was followed by overnight incubation. The initial 3 M urea treatment did not remove the small proteins as observed in Figure 12A. It is possible that the urea prepared in this experiment was lower than the urea in the previous experiment. The arrow between the 3rd and 4th lane from the left indicates a small shift of the gp270 band, possible because the 3 M urea triggered the proteolysis of gp270 in the capsid.



Reassociation of gp9 with the capsid in low concentration of urea

2.5.3 Guanidine chloride treatment of the PAU empty capsid

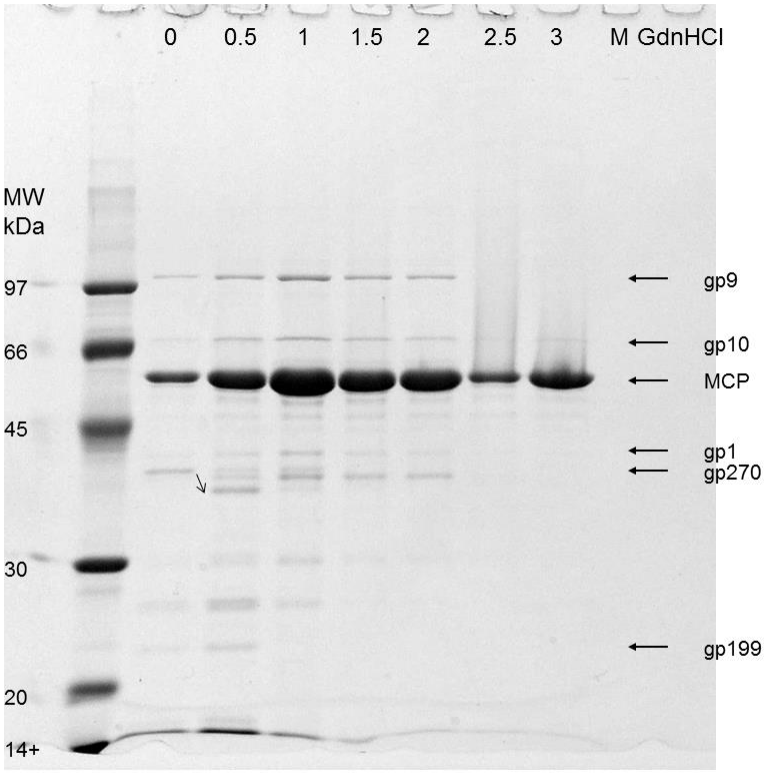
Unlike urea, which mildly weakens the hydrophobic interaction of protein, guanidine chloride (GdnHCl) is one of the strongest protein denaturants. It disrupts electrostatic interactions and hydrogen bonding of protein, causing the protein to lose ordered structure at high concentration (Lim, Rosgen et al. 2009). Because urea and GdnHCl target different interactions in protein denaturing, the stability of a protein measured by urea and GdnHCl denaturations may lead to different conclusions, depending on the importance of hydrophobic interactions and electrostatic interactions to the protein (Monera, Kay et al. 1994). The PAU empty capsid was treated with GdnHCl with a protocol similar to the urea treatment experiment.

The SDS gel of the GdnHCl-treated capsids shows that the PAU capsid is stable in 2 M or less GdnHCl (Figure 14). With 1 M or more GdnHCl, bands of the small proteins disappear from the gel, but the bands of gp1, gp9 and gp10 maintain their intensity relative to the MCP band. Therefore, the small proteins left the PAU capsid treated with 1 - 2 M GdnHCl, but gp1, gp9 and gp10 were not affected. With 2.5 M or more GdnHCl, the MCP bands are weaker and smearing on the gel, suggesting that the recovery of capsid protein through ultracentrifugation was reduced and the protein might be oxidized during the experiment. So it is likely that with 2.5 M or more GdnHCl some capsids were broken, and the MCP lost its native structure, which made the hypothesized oxidation possible. The PAU capsid's instability in GdnHCl and the stability in urea suggest that electrostatic interaction or salt bridging may be the main interactions to hold the capsid together. The small proteins may loosely interact with capsid shell with both hydrophobic interaction and electrostatic interaction, so they are susceptible to both urea and GdnHCl.

The gel shows an odd apparent shift of the gp270 band by about 2 kDa at 0.5 M GdnHCl, a concentration at which the other small proteins remained in the capsid. A similar band position shift can also be observed in the capsid treated with 1.5 M urea (Figure 13). It could be that at a low concentration of urea or GdnHCl, the structure of gp270 was relaxed and became accessible to the still functional protease gp151.

Figure 14 SDS gel of PAU empty capsids treated with GdnHCl

PAU empty capsids were treated with different concentrations of GdnHCl, labeled on the top of the gel. Some capsid protein bands were pointed by arrows and texts on right. The arrow on the left (between lane 2 and lane 3 from left to right) indicates an about 2 kDa shift of the position of the gp270 band.



SDS gel of PAU empty capsids treated with GdnHCl

2.5.4 Heating the PAU empty capsid

Surface decoration by minor capsid protein is a common feature in phages, especially in mid-sized and jumbo phages. The decoration density is thought to stabilize capsid structure in certain conditions but this role has rarely been tested. As a follow-up to urea-extraction of the surface decoration protein gp9, we were interested to learn if losing gp9 changes the thermal stability of the capsid. PAU capsids that lost gp9 were incubated in a series of temperatures, pelleted by ultracentrifugation and then checked by gel. To see if gp9 contributes to the thermal stability of the capsid, an assumption must be held that gp9 is stable on the capsid before the capsid falls apart., This is not the case for the PAU capsid (Figure 15A), arguing that the main function of gp9 is not to stabilize the capsid. Temperature is actually another condition that can be used to manipulate the minor capsid proteins of the PAU capsid.

The PAU capsid is stable at temperatures below 76˚C. EM images of the heated PAU capsids display no significant difference compared to the control capsid (Figure 15C). However, the capsid could not survive 80˚C (Figure 15A). The agarose gel of capsids heated at a series of temperatures shows that capsids heated at higher temperature migrate faster in the gel (Figure 15A), similar to the capsids treated with urea. So it is likely that, starting at 50˚C, the capsid began to lose surface decoration proteins in a temperature dependent fashion. The SDS gel of the capsid heated at 70˚C shows that the heating reduced the amount of gp9 and gp10 protein in these capsids, but the amount of the other minor capsid proteins did not change (Figure 15B).

High temperature destabilizes protein structure by weakening hydrophobic interactions and hydrogen bond networks, and salt bridges have been repeatedly shown to enhance the thermal stability of protein (Folch, Dehouck et al. 2010). The dissociation of gp9 by heating is very similar to what has described in the urea extraction section, confirming that hydrophobic

interaction is the main interaction between gp9 and the capsid shell. The removal of gp10 by heating suggests that hydrogen binds may play an important role in the interaction between the protein and the capsid. Surprisingly, all the small proteins that were easily removed by urea or GdnHCl were stable in the capsid at 70˚C. These proteins are rich in charged amino acid residues and a few of them have quite high pIs. Lots of salt bridges may be formed to hold all the small proteins together, and also maintain the interaction between these proteins and the inner capsid surface which was shown to be negatively charged in the HK97 prohead II (Conway, Wikoff et al. 2001).

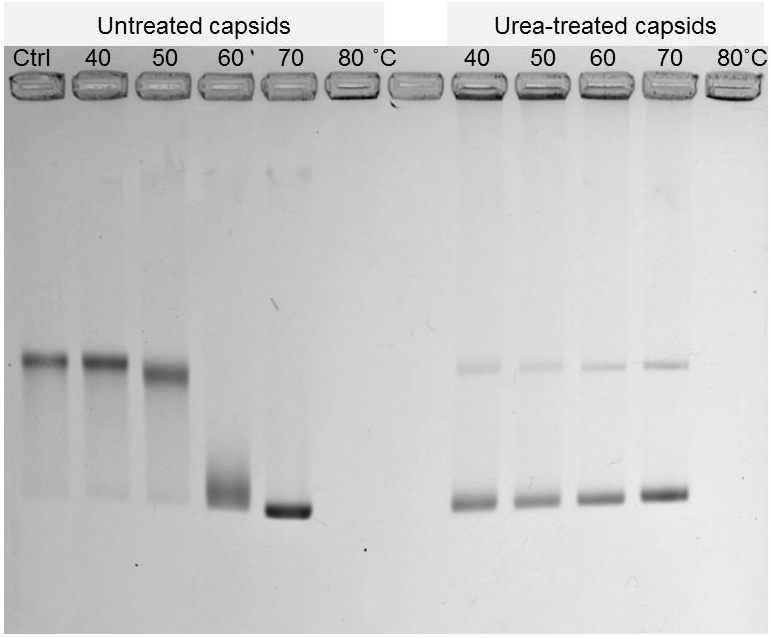
Figure 15 Heated PAU capsids

A. Agarose gel of PAU empty capsids heated at a series of temperatures. Capsids with or without treatment by 6M urea were heated at 40, 50, 60, 70 and 80˚C. No bands were formed for capsids treated at 80˚C. Bands of capsids without urea treatment shift lower when the temperature was higher than 50˚C, while bands of urea-treated capsids show no significant difference at different temperature. However, a faint band at the same position as the control remains for capsids treated with urea.

B. Agarose gel and SDS gel of capsids heated at 70˚C. A new batch of slightly less purified capsids were used in this experiment, so the non-capsid protein bands (such as the band below gp270) are stronger than the same band in the previous SDS gels. Ctrl2 is the old, more purified sample. The gel shows that the gp9 and gp10 bands become weaker for the heated sample. But the other bands show no significant differences.

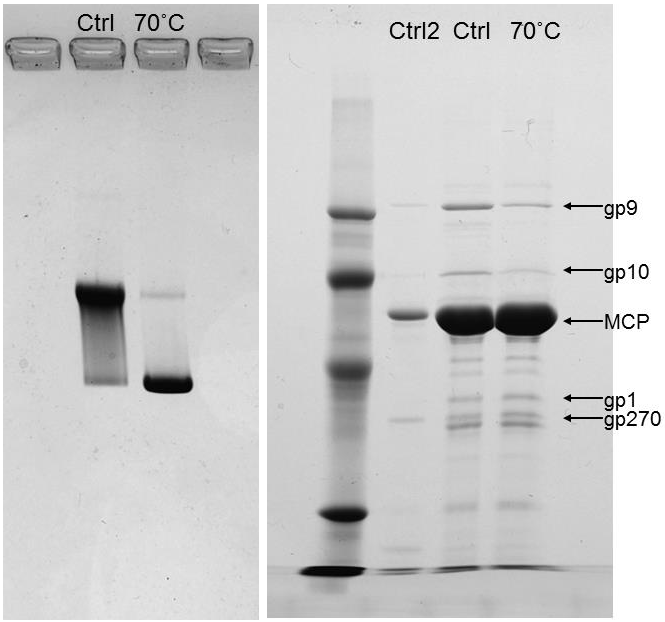
C. Negative stain EM images of heated capsids. The heated capsids display no significant difference to the control capsids.

A



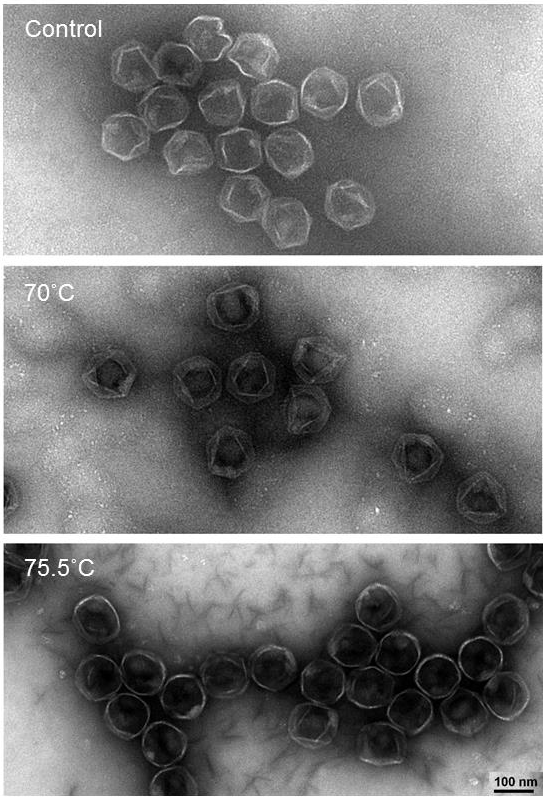
Agarose gel of PAU capsids heated at a series of temperatures

B



Agarose gel and SDS gel of PAU capsids heated at 70˚C

C



Negative stain EM pictures of heated capsids

2.5.5 EM structure of capsids treated with urea

The sequencing data and the chemical manipulation of the minor capsid proteins suggest that gp9 is a surface decoration protein that forms the banana density and the small proteolytically processed proteins are internal proteins, which leave gp1 and gp10 as the likely candidate proteins for the mango density under each vertex. Again by collaboration with Dr. James Conway and Dr. Katerina Toropova, we obtained an 11.5Å density map of the PAU empty capsid treated with 7M urea. The density map together with cryo-EM micrograph provides conclusive evidence on the location of the minor capsid proteins.

The cryo-EM image of PAU capsids shows that each capsid has an irregular patch inside the capsid (Figure 16A, left). This density tends to be located in a corner inside the capsid, so it may loosely interact with the portal complex or one of the eleven vertices. Interestingly, the irregular density is completely missing in the cryo-EM image of the capsids treated with 7M urea (Figure 16A, right). Because the capsid loses gp9 and the small proteins at 7M urea, and evidence suggests that gp9 is a surface decoration protein, the small processed proteins that leave the capsid in presence of 3M or more urea are the obvious candidates for forming the internal patch.

The cryo-EM model of the capsid treated with 7M urea has no banana density around the pentamer, but otherwise is very similar to the 9Å model of the untreated capsid (Figure 16B). Fine differences in the surface details between the two models can be observed, a result of the resolution difference which disappears when the 9Å model was calculated to 11.5Å (Figure 16C). Therefore, the shell structure was not detectably affected by the urea-treatment. Without gp9, the hexamer around the pentamer displays local 6-fold symmetry and is very similar to other hexamers, so gp9 is the sole protein of the banana density. The mango density remains on

the interior of the capsid shell, providing strong evidence that the gp1 and gp10 remaining after 7M urea are the mango proteins.

Unfortunately, the resolution of the urea-treated capsid map is too low to trace the long rod-like density observed in the 9Å map. As a result I am not able to show if losing the banana density results in significant changes in the conformation of the hexamer near the pentamer.

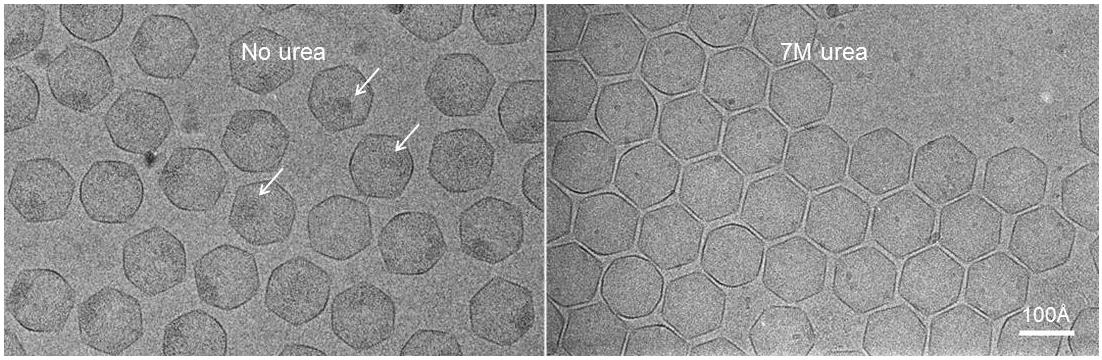
Figure 16 Cryo-EM micrographs and structure of the urea treated PAU capsids

A. Cryo-EM images of capsids with and without 7M urea treatment. Arrows on the left point to the irregular dark patch inside the capsids.

B. Cyro-EM model of the capsid treated with 7M urea. The resolution of the model is 11.5Å, so the detail of the surface is slightly different from the surface of the 9Å model in Figure 8. The black circle indicates the approximate area that is compared below.

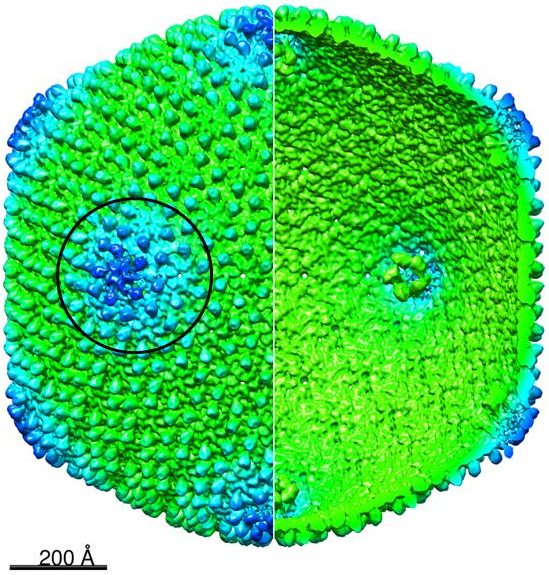
C. The surface structure around the pentamer of capsids with and without 7M urea treatment. The banana density is colored in yellow. The pentamer density is colored in purple. The other density is colored in blue. Courtesy by Dr. James Conway and Dr. Katerina Toropova.

A



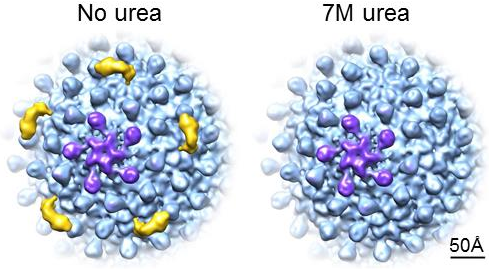
Cryo-EM images of PAU empty capsids treated with/without 7M urea

B



Outer and inner surface of the PAU empty capsid treated with 7M urea

C



Comparison of structure around the pentamer for capsids with/without 7M urea

2.6 DISCUSSION

2.6.1 A phage distant to others

PAU is likely a phylogenetically isolated phage that has diverged a considerable distance in evolution from other phages we are familiar with. Homology search of gene products of potential PAU ORFs by blastp suggests that the genome sequence of PAU is very distant from other sequenced phages. As many fewer jumbo phages have been isolated and sequenced compared to the small or mid-sized phages, it is not surprising to sequence a new jumbo phage and find that many gene products exhibit no sequence homology to proteins with known function. For example, when the genome sequence of the first T=27 phage phiKZ of *Pseudomonas aeruginosa* was published, 59 gene products showed obvious homology to proteins in the GeneBank database at that time (Mesyanzhinov, Robben et al. 2002). However, greater sequence homology to phiKZ gene products were discovered later when more large *Pseudomonas* phages with similar size to phiKZ were sequenced (Cornelissen, Hardies et al. 2012). Therefore, PAU may be the first member of a group of jumbo phages that is yet to be discovered.

The PAU data show novel characteristics in capsid structure and tail morphology. It is the only phage known that has a T=25 capsid. The location of the banana and mango density is also unseen in other virus capsid structures. It is not clear if this type of decoration density is dedicated to T=25 capsids, as no additional T=25 phage has been discovered yet. The morphology of the PAU baseplate also brings more “personality” to the structure of the PAU virion. The narrow appearance of the PAU baseplate does not resemble the T4 baseplate model that all other phages with contractile tails seem to follow (Leiman, Arisaka et al. 2010).

A general explanation for the novelty of a new jumbo phage like PAU lies in the fact that very few jumbo phages have been isolated and studied in detail. Jumbo phages appear to be easily excluded in the traditional spot-on-plate method for phage discovery, dissociated by the use of CsCl gradients during purification, or ignored due to their slower growth compared to their small or mid-sized capsid peers. Another reason lies in the special environment that PAU was isolated from. Unlike most phages isolated from sewage, natural water or soil, PAU was isolated from diseased silkworms along with an even bigger phage that could not be grown (Ackermann, Auclair et al. 1994). The silkworm breeding bed is an artificial environment piled with silkworm excrement and mulberry leaf debris and therefore rich in microorganisms. PAU may be one member of a group of bacteriophages that has recently emerged and has been quickly evolving in this environment.

2.6.2 A novel T-number formed by a familiar fold

The "quasi-equivalence theory" developed by Caspar and Klug in 1962 provided elegant conceptual fundamentals to understand the organization of icosahedral capsids (Caspar and Klug 1962). According to this theory, any T-number ($T = h^2 + hk + k^2$, where h and k are integers) is allowed for construction of an icosahedral capsid. However, most isolated tailed phages are small and mid-sized, leaving many possible T-numbers unobserved in phage capsids. An important goal of my jumbo phage project is to find the missing T-numbers from our jumbo phage collection. The T=25 phage PAU is one of the five jumbo phages with novel T-numbers, filling empty slots in the table of all possible T-numbers by the quasi-equivalent formation of icosahedral capsids.

The HK97 fold serves as the main building element for icosahedral capsids of bacterial and archaeal tailed phages and mammalian herpesviruses, and for spherical particle encapsulins in bacteria and archaea (Suhanovsky and Teschke 2015). Evidence of the shared fold for proteins are clear at the 3D structural level, but similarity in amino acid sequence is usually low. Though the primary sequence of gp156 shows little sequence identity to other capsid proteins, the 3D structure prediction of gp156 and the cryo-EM structure the PAU capsid reveal that gp156 adopts the HK97 fold, adding another example of an icosahedral capsid protein sharing the HK97 fold. The major capsid protein of two other jumbo phages studied in our lab, the T=28 121Q and the T=52 G, again have the HK97 fold. Therefore, the HK97 fold is apparently ubiquitous in tailed phages and jumbo phages are no exception regardless of their sizes.

The 9 Å density map of the PAU capsid shows some notable variations in subunit conformations. Two subunits in the hexamer near the pentamer have observable differences in 3D structure compared with the other hexamer subunits, including subunit B which the banana density makes the most contact with, and subunit A which is the main interacting subunit to the pentamer. The conformational deviation of subunit B is more evident because superimposing the density of hexamer #2 with hexamer #1 shows that the long rod-like density in subunit B of hexamer #2 is about 10 Å lower than in the other subunit. Although in theory the quasi-equivalent packing of MCP subunits in the asymmetric subunit should result in different subunit conformations, significant structural variations in the backbone of HK97 fold in hexamer subunits are not observed in other mature icosahedral capsids. The hexamer in the mature HK97 capsid displays strong 6-fold symmetry and the structural difference is limited to the flexible ends of the N-arm and E-loop (Wikoff, Liljas et al. 2000). In another T=7 phage ε15, the near atomic resolution cryo-EM model also shows that the variations in the hexamer are limited to the

N-arm and E-loop (Jiang, Baker et al. 2008). Therefore, the presence of the banana protein, which is unique to PAU, is the likely cause of conformational deviation of subunit B from other hexamer subunits.

In addition to the core HK97 fold that builds the continuous capsid shell, the N-terminal region of gp156* forms the knob-like accessory structure on the outer surface. The location of this accessory structure on the contacting interface of two nearby subunits in the same capsomer suggests that it may strengthen the intracapsomer interaction by increasing the contacts with a neighboring subunit. Many capsid proteins were reported to contain extra domains which come with various size, structure and location but usually with a common function of stabilizing the capsomer. For example, the T4 vertex protein gp24 has a 60 AAs insertion domain in its E-loop, a chitin-binding-like domain that interacts with a neighboring subunit in the same pentamer (Fokine, Leiman et al. 2005). A group of P22-like short tailed phages (P22, Sf6 and CUS-3) have a conserved extra external domain with about 120 AAs (Parent, Gilcrease et al. 2012). This external domain forms a surface protrusion on the capsid, similar to the knob-like structure feature in PAU. Analysis based on the cryo-EM reconstruction of the three phages suggests this domain is involved in intercapsomer interaction across a local two-fold symmetry (P22 and CUS-3) or intrahexamer interaction with neighboring external domains (Sf6) (Parent, Gilcrease et al. 2012).

2.6.3 A unique arrangement of minor capsid proteins

Ten capsid proteins in addition to the MCP gp156 were identified by N-terminal sequencing, and assigned to the structural features of the empty PAU capsid by comparing the SDS gel and the cyro-EM structure of the urea treated and untreated capsids. Minor capsid proteins and internal

proteins are commonly found in large bacteriophages. For example, the T4 mature capsid contains soc and hoc on its outer surface, and IP I*, II*, III* and alt* inside the capsid (Black 1974; Bijlenga, Ishii et al. 1978; Goff 1979; Kaliman, Khasanova et al. 1990). The T=27 phage phiKZ contains several proteins with high copy numbers per capsid that may contribute to the "inner body" structure (Thomas, Weintraub et al. 2012; Wu, Thomas et al. 2012). The structure of another T=27 phage phiRSL1 suggests that the phiRSL1 capsid may have several decorating proteins on its surface (Effantin, Hamasaki et al. 2013). The abundance and high sequence divergence of jumbo phage structural proteins combined with poor resolution of SDS-PAGE of whole phage samples present a particular challenge in studying the capsid of a newly isolated jumbo phage. The lack of sequence homologs hinders the precise localization of those proteins in the capsid of jumbo phages in which the organization of capsid proteins is diverse and poorly understood. For example, the structure and protein of phiKZ have been studied for over a decade, but the identities of its decoration proteins are still unknown. The sequential dissociation of the internal protein and the decorating protein using denaturing reagents produces empty PAU capsids that contain reliable information about the location and function of those proteins. Since the mature empty capsid is a common by-product during phage preparation in a nutrient-rich laboratory setting, the use of such a mild denaturing reagent for extracting internal and decoration proteins from the empty capsid can be an effective strategy to study the structure and proteins of new jumbo phages.

In addition to the MCP that is conserved across all tailed phages, there is a wide variety of less conserved accessory proteins on the surface of many phage capsids. These additional capsid proteins are found in many small-sized phages and almost all mid-sized and jumbo phages, but are far less studied than the major capsid. The accessory capsid proteins are often named

decoration proteins because most of them dispensable in laboratory conditions, and are attached after procapsid assembly. I did an extensive search on tailed-phage decoration proteins that have been identified either by biochemistry and/or high resolution cryo-EM. These proteins can be arranged under three categories based on their location on the surface geometry:

1. Decoration proteins on the interface between capsomers. This is the most common location of decoration proteins. Many decoration proteins in this category are trimers adhering to the local three-fold between three capsomers, such as gpD in the T=7 phage λ (Sternberg and Weisberg 1977), dec in the T=7 phage L (Tang, Gilcrease et al. 2006), gp123 in the T=13 phage phi92 (Schwarzer, Buettner et al. 2012) and gp8.5 in the T=3 prolate phage phi29 (Morais, Choi et al. 2005). There are two other less common types of morphologies: soc in the T=13 prolate phage T4 forms continuous mesh encircling hexamers (Fokine, Chipman et al. 2004); and i and j in the T=7 marine phage Syn5 and gp29.2 in the T=16 phage SPO1 forms asymmetric monomeric structure near the local but not global 3-fold symmetry axes (Duda, Hendrix et al. 2006; Hua 2010; Gipson, Baker et al. 2014).
2. Decoration proteins on the center of hexons. Examples of this type of decoration protein include: hoc in T4 (Fokine, Chipman et al. 2004), pb10 in the T=13 phage T5 (Effantin, Boulanger et al. 2006), h in Syn5 (Gipson, Baker et al. 2014) and a mushroom-shaped unidentified structure in the T=12 marine phage SIO-2 (Lander, Baudoux et al. 2012).
3. Decoration proteins on pentons. This type includes the turret-like molecule above the pentamer in the T=7 Great Salt Lake phage CW02 (Shen, Domek et al. 2012)

and the extra structure on the pentamer in SIO-2 (Lander, Baudoux et al. 2012). (It is reported that turret shaped density at the vertex of the T=16 marine phage Syn9 may be a decoration protein (Weigele, Pope et al. 2007), but I incline to believe that it is a vertex protein like gp24 in T4 because there is no capsid shell density bellow the turret density.)

An intriguing observation about the banana and mango proteins is that neither of them is arranged in any location listed above. The banana protein mainly interacts with subunit B in hexamers surrounding the pentamers, bridging between the center of the hexamer and the protruding accessory domain of subunit B. The mango proteins are unconventionally attached to the pentamer in the interior of the capsid and form a pentameric structure. In spite of their unique appearance and location, the banana and mango proteins seem likely to serve the same purpose as the other accessory capsid proteins of stabilizing the capsid shell. One possible reason for such an extraordinary arrangement of accessory proteins is that the major capsid protein of PAU may be optimized for constructing flat hexamers which form a less curved capsid shell, resulting in a much weaker area around the penton vertex where the capsid shell is more curved. The presence of the banana and mango proteins at the pentamer and its near hexamers may solve this problem. In fact, the banana and mango proteins may do a great job in stabilizing the PAU capsid, because PAU is the only jumbo phage that can be purified by CsCl gradients. Phages 121Q, PBS1 and N3 are unstable in CsCl.

The banana and mango proteins may be incorporated into the capsid at different stages of capsid assembly. The banana density in the cryo-EM model is more blurry than the capsid shell. The banana protein leaves the capsid at 5 M or higher urea, and this process seems partially reversible. So the banana protein is probably a decoration protein that is attached after capsid

maturation. The mango density is as rigid as the capsid shell and could never be extracted by urea or guanidine chloride without disrupting the capsid shell. Therefore, the mango proteins are likely attached to the pentamer during procapsid assembly.

Another interesting obervation about the banana and mango proteins is that the three genes encoding these accessory capsid proteins are nearby each other in the genome. In contrast, the other structural genes, including the classic portal-protease-scaffolding-MCP cluster, are separated by other genes. Possibly, the banana and mango genes may have been acquired by PAU recently in the phage's evolution. This horizontal gene transfer, if it is true, is probably coupled with an expansion of the capsid size from a smaller T-number to T=25, considering their important roles in capsid structure.

I also identified seven small internal proteins in the PAU empty capsids. These proteins form the irregular patch visible under cryo-EM microscope, which loosely interacts with the portal complex or a pentamer and can be extracted with low concentrations of urea. However, they do not form an "inner-body" structure, a long protein core sensitive to high electron dosage during cryo-EM imaging, found in phiKZ (Wu, Thomas et al. 2012) or 121Q (Chapter 3). These proteins are small, proteolytically processed at the N-terminal end at Glu-Ala sites, and some of them are highly positively charged. The similar characteristics to the IPs in phage T4 suggest that the internal proteins may be injected into the host and bind the negatively charged PAU DNA for protection against host restriction enzymes.

3.0 THE CAPSID STRUCTURE AND PROTEIN OF COLIPHAGE 121Q

3.1 INTRODUCTION

Phage 121Q, which resembles the type 121 phage found in Europe, was isolated from sewage in Quebec, Canada by Achermann and Nguyen in 1982 (Ackermann and Nguyen 1983). For 35 *Escherichia coli* strains enriched from sewage, 121Q was only active on an uncommon *Escherichia coli* strain named MulB 70.1. Our previous study on the cryo-EM structure of the empty phage shows that 121Q has a T=28 capsid with a very rough surface. The genome sequence 121Q shows high similarity to several coliphages with similar genome sizes isolated from sewage in Europe and Asia, but very poor similarity to sequences of other organism. Therefore, 121Q may belong to a group of jumbo phages that infect sewage bacteria all over the world.

We are interested in 121Q for two reasons: that it is the first T=28 phage discovered and that it infects *Escherichia coli* which may allow genetic manipulation of the jumbo phage in the future. This chapter covers the purification of 121Q particles, cyro-EM structure of 121Q phages and identification of structural proteins by mass spectrometry.

3.2 PURIFICATION OF 121Q PARTICLES

3.2.1 Morphology of 121Q

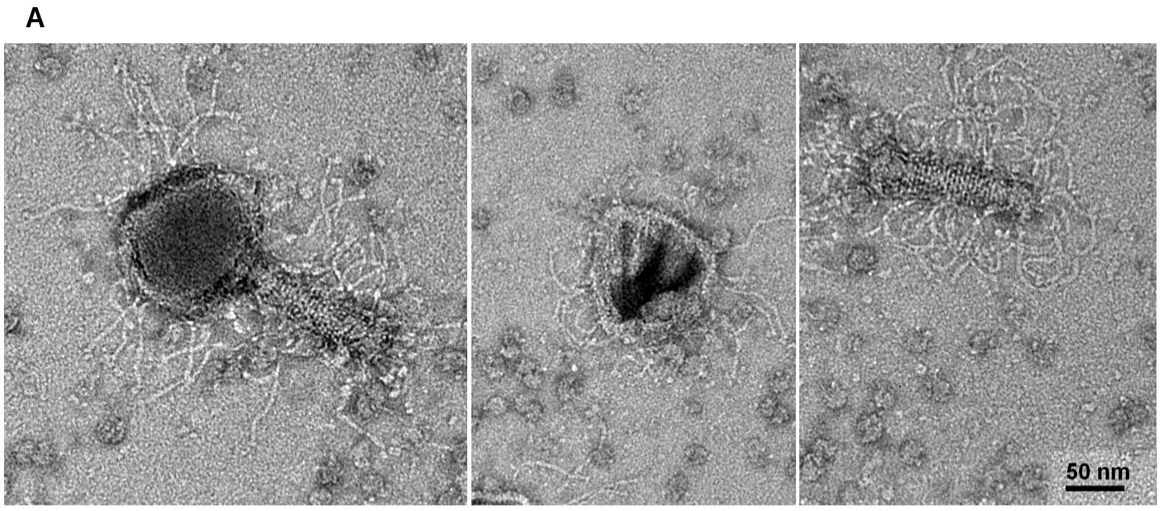
The capsid of 121Q is icosahedral, connected to a contractile tail by a narrow neck (Figure 17). The diameter of the capsid is 140 nm along the axis of 5-fold symmetry and 120 nm along the axis of 3-fold symmetry. The tail of 121Q is relatively shorter than other jumbo phages, measuring about 131 nm including the neck and the baseplate. DNA is compressed and organized inside the capsid, forming an evenly spaced grid in the cryo-EM images. However, the DNA grid is not observable near the portal axis, probably because the presence of the 'inner body' (described below) interferes with the arrangement of DNA (3.3.4).

121Q is a 'hairy' phage, with hair-like fibers stretching from both the capsid and the tail. The length of both the capsid hair and the tail hair is approximately 100 nm. The hairs appear attached to multiple sites on the surface of the capsid and the tail sheath. The number of hairs that are attached to the capsid or to the tail sheath is not fixed, ranging from 0 to about 20 by estimating from EM images. The phage might slowly lose the hair in laboratory conditions because phages that had been stored in the refrigerator for over a year had much less hair. Dark spots around the distal end of the hair can be observed in the cryo-EM image, suggesting that the tip of the hair may contain heavy atoms or a globular domain that absorbs the electron beam really well (Figure 17B). Long hair on both the capsid and the tail sheath has not been reported for other phages. The hair may be involved in interacting with the host membrane because it is very difficult to separate 121Q particles from host cell debris during purification.

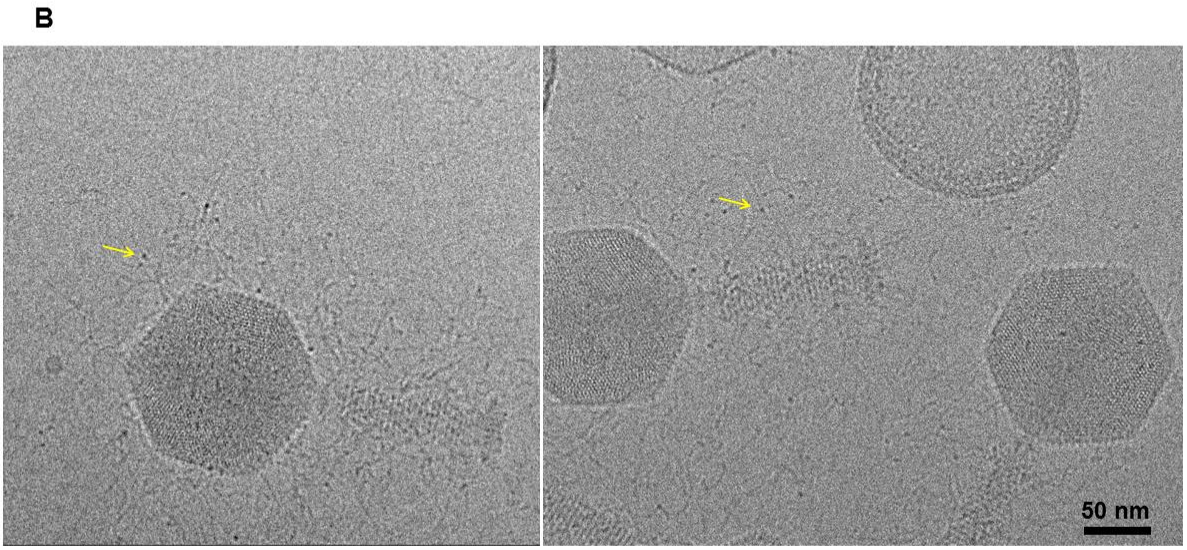
Figure 17 Electronic microscopy images of 121Q phages and parts.

A. Negative stain EM images of a 121Q phage, an empty capsid and a tail. All three particles have hairs attached.

B. Cryo-EM images of 121Q phages. Arrows point to the dark dots at the distal end of the hair. The compressed DNA is organized, showing grids in the micrograph, except near the axis from the portal to the distal vertex.



Negative stain EM images of a121q phage, an empty capsid and a tail



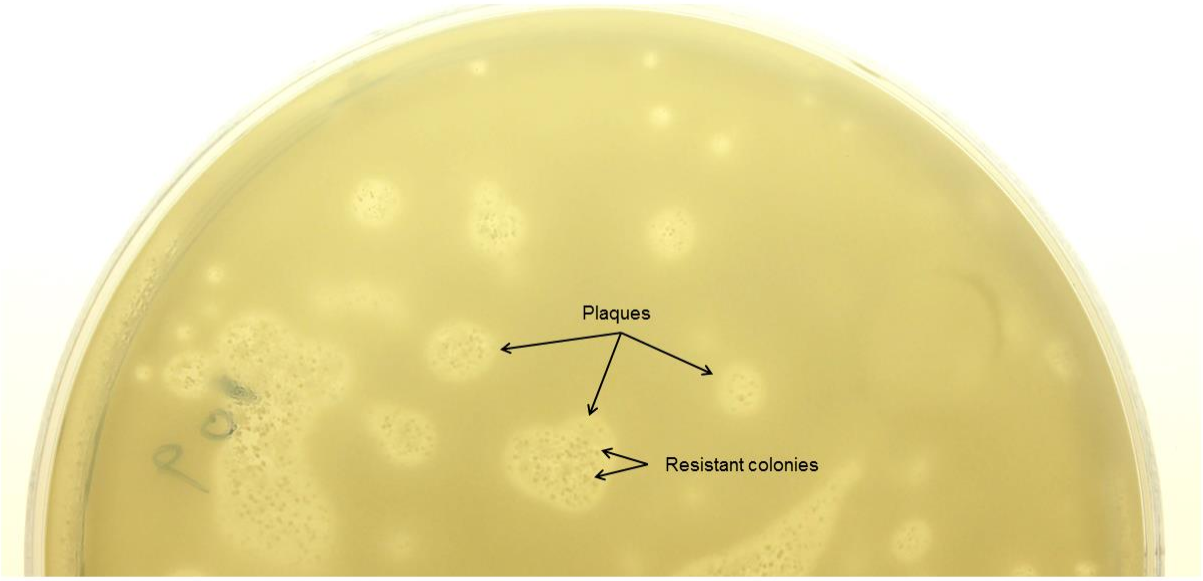
Cryo-EM images of 121Q phages

3.2.2 Growth of 121Q

High-titer 121Q stock was prepared using the plate stock method. 121Q cannot be grown in LB liquid culture but forms small plaques on a standard double-layered LB plate. When warm LB with 0.2% agarose was used as the top layer, the phage mixed with fresh overnight culture of *Escherichia coli* MulB 70.1 formed large plaques with small colonies resistant to phage infection. The resistant colony becomes visible after the plate was incubated in 37 ℃ for 14 hours, so the incubation in warm room should not exceed 16 hour. The quick occurrence of the resistant host colony suggests that the 121Q receptor on the MulB 70.1 cell surface may be dispensable. We obtained about 30 ml white-colored (light scattering) $1X10^{11\sim12}$ pfu/ml stock from plates containing 1 L LB agar.

Figure 18 A picture of a 10 X 10 cm plate of 121Q.

Arrows point to plaques and resistant colonies in a plaque.



A 10 X 10 cm 121Q plate

3.2.3 Purification of 121Q particles

The sensitivity to CsCl and the high binding affinity to host membrane make the purification of 121Q particles difficult. The phage quickly releases DNA when it is treated with 1.2 g/mL or higher CsCl solution, forming no band in overnight equilibrium gradients or poorly separated bands with high viscosity in step gradients. Consequently, we had to use sucrose gradients which were less effective in separating phage particles from the cell debris than CsCl gradients. Running the 121Q lysate through 15~45% continuous sucrose gradients generated tail, empty capsid and phage bands with lipid contamination (Figure 19A). 121Q particles showed high affinity to host membrane in the purification process. Repeated sucrose gradients combined with treatments using detergents, trypsin and DEAE filtering could not completely remove the membrane contamination from the tail and the empty capsid samples (Figure 19C). Because both the tail and the capsid interact with the membrane, the final tail sample contains a small amount of capsids and the capsid sample contains a small amount of tails.

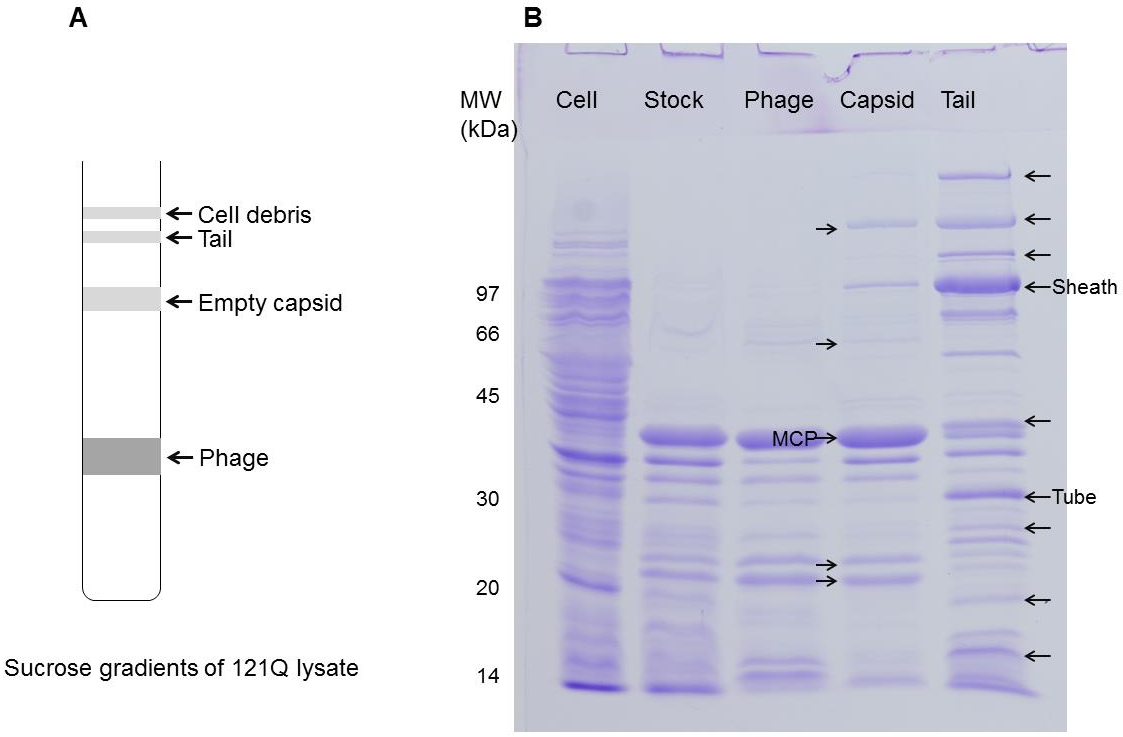
The SDS-PAGE gel of the 121Q particles shows that both the tail and the capsid contain lots of structural proteins (Figure 19B). The heavy bands can be quickly categorized as capsid or tail proteins, but the poor purity of the samples make it difficult to identify the less abundant tail or capsid proteins. The identification of tail and capsid proteins is discussed in detail in 3.4.

Figure 19 121Q sample purification

A. The pattern of the initial sucrose gradients of concentrated 121Q lysate. Four bands are formed in the 15-45% continuous sucrose gradients. Contents of the bands from top to bottom are: cell debris, tails, empty capsids and phages.

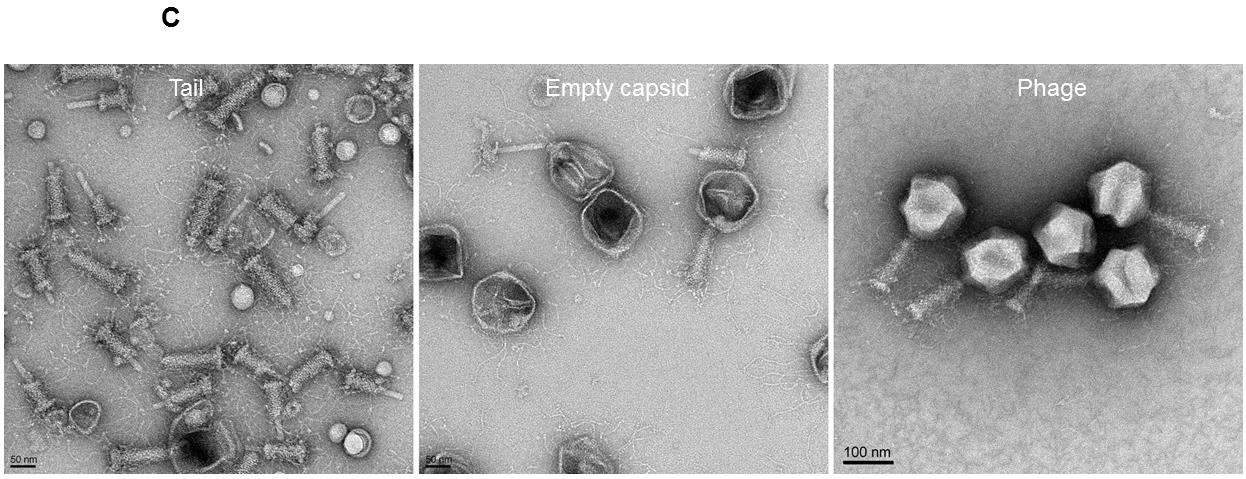
B. SDS gel of 121Q samples. The tail, empty capsid and phage samples were purified by three rounds of sucrose gradients. The host cell and lysate stock samples are controls. Arrows indicate possible capsid or tail proteins. I did not do any band identification.

C. Negative stain EM images of 121Q samples. The phage sample is quite pure, but the tail sample contains some cell debris (the small particles). The empty capsid sample contains a small amount of host debris. Contamination of tails or empty capsids is found in the empty capsid sample or the tail sample.



Sucrose gradients of 121Q lysate

SDS gel of 121Q samples purified by gradients



Negative stain EM images of 121Q samples purified by gradients

3.3 STRUCTURE OF 121Q PHAGES

3.3.1 Predicted structure of the 121Q major capsid protein

The MCP of 121Q was identified by BLAST as gp585, a 42.2 kDa precursor protein with 391 AAs before proteolysis. The exact position where the N-terminal is cleaved off from the precursor has not been determined by N-terminal sequencing. However, our mass spectrometry data of mature particles show heavy coverage starting at R58 but not a single peptide covers the first 57 AAs, so it is very like that less than 56 AAs are cleaved off by the protease. Since capsid protein processing mostly happens specifically at the Glu-Ala site, the first Glutamic acid residue upstream of R58, E43, is likely where the protease cleaves gp585.

Homology search by PSI-BLAST shows that gp585 has sequence homology to (putative) MCPs in many phages. However, sequence homology to well-studied MCPs is low, so modeling the secondary structure may be a better way to provide some insight into the gp585 structure. The secondary and 3D structure modeling performed by Phyre2 (Kelley and Sternberg 2009) shows high percentage of alignment to the 427 AAs T4 vertex protein gp24 (PDB ID: 1yue (Fokine, Chipman et al. 2004)), matching residues 46-374 of gp585 to residues 8-392 of the gp24 (Figure 20). Residues 145-374 of gp585 align with residues 41-62 and 156-392 of gp24 which form the two core domains of the HK97 fold, the A- and P-domains.

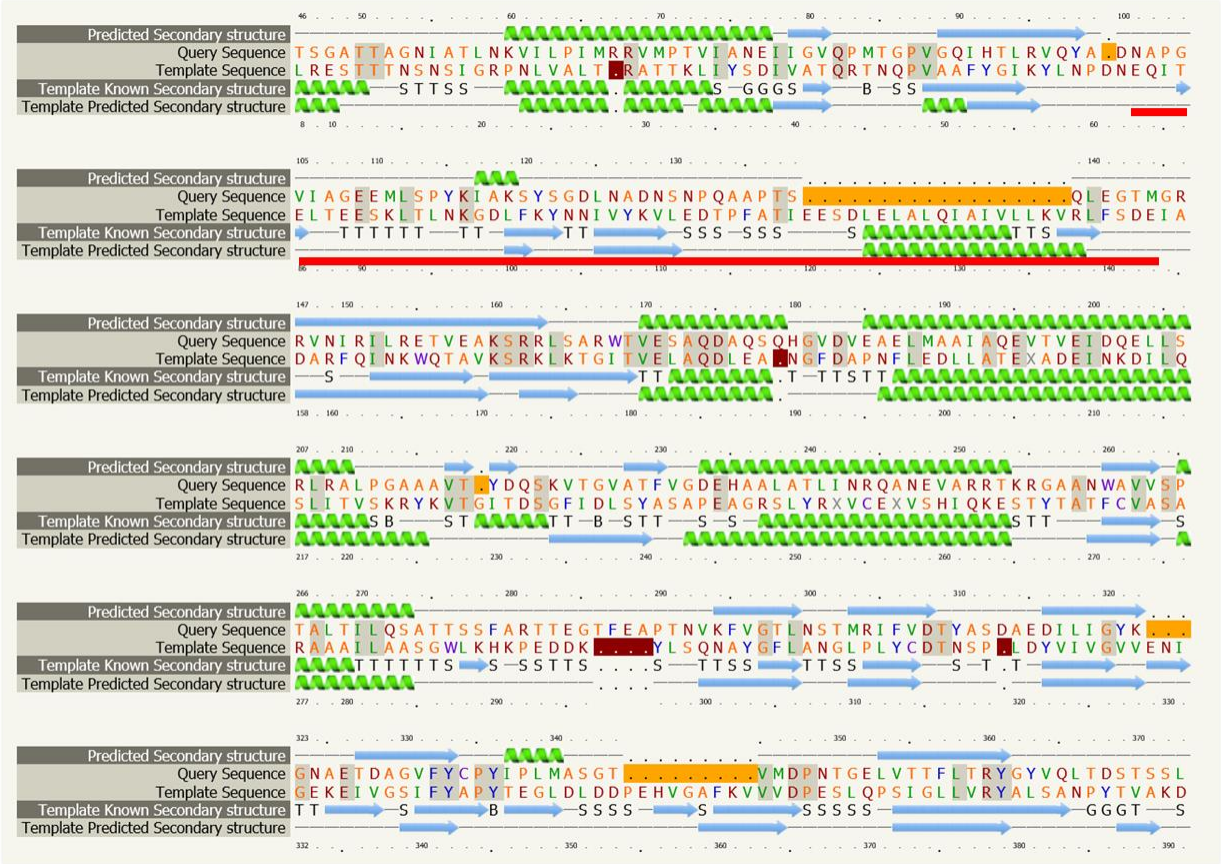
Residues 46-144 of gp585 align with residues 8-40 and 63-155 of gp24. The alignment is consistent with the prediction of cleavage at E43 based on the mass spectrometry data. Residues 2-40 of T4 gp24 form a compact helix-turn-helix structure that leans against the signature long helix and the β-sheets of the P-domain, compared to the extended N-arm in the HK97 MCP structure. Residues 63-155 of gp24 forms the insertion domain and two flexible linker regions

that extend this domain to the surface of the counterclockwise gp24 subunit, compared to the E-loop in the HK97 MCP structure. A gap of 18 residues is formed in the gp585 sequence, corresponding to residues 120-137 of gp24 which forms the C-terminus of the insertion domain. Therefore, if gp585 has an insertion domain as gp24 does, it will be a smaller one.

The 121Q MCP gp585 undoubtedly adopts the HK97 fold. The coverage of the alignment to T4 gp24 suggests that gp585 is more close to gp24 in evolution than the PAU MCP which only shows alignment coverage in the region that forms the core HK97 fold. While the N-terminal region of gp585 does not necessarily form the insertion domain and the compact helix-turn-helix structure, the Phyre model suggests that gp585 may not have any domain larger than the insertion domain in addition to the HK97 core. When analyzing the cryo-EM structure of the 121Q capsid, any additional large chunk of density that is far away from the backbone of the HK97 fold is probably formed by a decoration protein but not by the N-terminal of gp585.

Figure 20 Secondary structure prediction of proteolytically processed gp156 by Phyre2.

The query sequence is the whole gp585 sequence. The template sequence is the T4 gp24 sequence. From top to bottom, the three lanes of secondary structure are the predicted secondary structure of gp585, the known secondary structure of gp24 (PDB ID: 1yue) and the crude predicted secondary structure of gp24 replaced with matched gp585 sequence. Green helices indicate α-helices and blue arrows indicate β-strands. The sequence of the insertion domain in gp24 is indicated by a thick red line. The last matched amino acid at the C-terminus is cut off in the figure.



3.3.2 Cryo-EM structure of the 121Q capsid

Cryo-EM models of 121Q were obtained by collaboration with Dr. James Conway and Dr. Alexis Huet in the Department of Structural Biology, University of Pittsburgh. We mainly worked on three models:

1. A 7.7Å model of the 121Q phage with icosahedral symmetry applied during 3D reconstruction. This model allows us to learn how the HK97 fold fits with the shell density and to find decoration density which is far away from the HK97 fold.
2. An 11Å capsid model without symmetry applied. This model provides the structure of the portal complex and the internal protein core.
3. Tail structures with 6-fold symmetry applied.

The pictures of the 121Q cryo-EM structure in this section are courtesy from Dr. Alexis Huet.

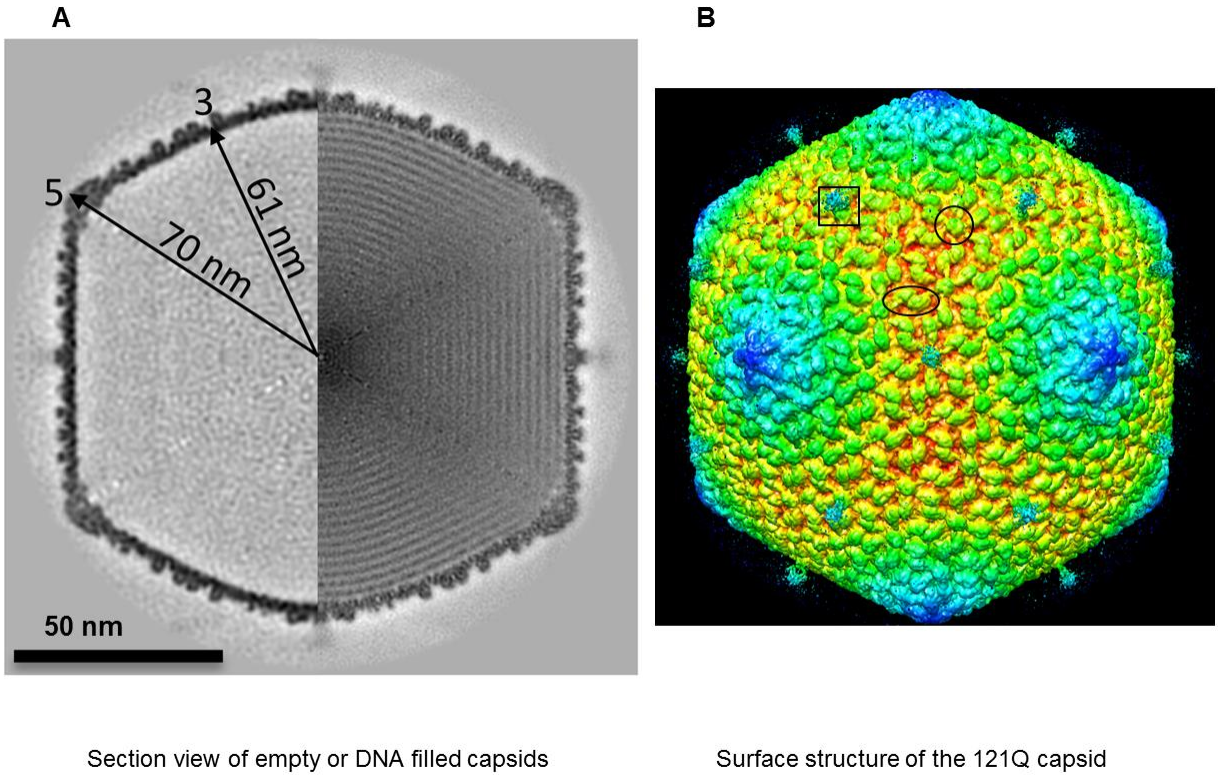
Cryo-EM models of the mature empty capsid without DNA and the phage capsid filled with DNA were reconstructed. The two models show very similar features in the section view of the capsid shell, except that the model of the empty capsid is slightly larger than the model of the DNA-filled capsid (Figure 21A). The model of the DNA-filled capsid has higher resolution and therefore was used for further analysis of the MCP fold and the decoration density. The model of the DNA-filled capsid shows a vertex-to-vertex diameter of 140 nm and a very rugged outer surface (Figure 21B). The averaged cryo-EM model of the T=28 (h = 4, k = 2, $T = h^2 + h*k + k^2 = 28$) capsid contains 270 hexamers and 12 pentamers formed by 1680 copies of MCP subunits. A chunk of protruding density is located on each subunit, forming a rotational symmetric pair near the local 2-fold symmetry of the two interacting capsomers. The pair of these densities interact with each other near the local 2-fold symmetry, and slightly extend to the surface of the

neighboring capsomer. The dimeric density does not reach the center of the capsomer. At the center of every hexamer sits a knob shaped density that does not contact with the dimeric density. Additional flexible density is located above the knob shaped density in the middle of the edge which is also the global 2-fold symmetry. This flexible structure may be related to the hair all around on the capsid in Figure 17.

Figure 21 Cryo-EM structure of the 121Q capsid.

A. Section views of two cryo-EM maps of mature capsids with or without DNA. The model of the capsid filled with DNA has a higher resolution and shows more structural details. The empty capsid is slightly bigger.

B. Cryo-EM model of the 121Q capsid. The surface is colored by radii to the center of the icosahedral capsid. Three types of potential decoration densities are highlighted by shapes. The dimeric density near the local 2-fold symmetry is enclosed by an ellipse. The knob shaped density at the center of the hexamer is enclosed by a circle. The flexible density at the global 2-fold symmetry is enclosed by a square.



3.3.3 Capsid density of the major capsid protein and the decoration proteins

Fitting the X-ray structure of the HK97 MCP gp5 (PDB ID: 1OHG) into the shell of the capsid model suggests the 121Q MCP gp585 adopts the conserved HK97 fold. Fitting the gp5 coordinates in one subunit in a hexamer shows that the A-domain and P-domain are enclosed by 121Q density that outline the secondary structure of the potential HK97 fold in the 121Q major capsid protein gp585 (Figure 22A). A long rod-like density fits well with the distinctive 40 Å long helix in the P-domain of gp5. The E-loop and N-arm of gp5 do not show good fits with the density map, either because the E-loop and N-arm are flexible or because gp585 might instead form structures that resemble the helix-turn-helix motif and the insertion domain in T4 gp24. Interestingly, although the density of the HK97 fold of gp585 forms the majority of the capsid shell, it does not cover the entire continuous and seamless capsid shell (Figure 22B and Figure 23B). Holes in the center of hexamers and between pairs of neighboring capsomers are observed when only the density of the HK97 fold is visualized. The round-shaped hole in the center of hexamers is about 30 Å in diameter. The dumbbell-shaped hole between neighboring capsomers is about 40 Å long and 20 Å wide at the two ends and 10 Å wide in the middle. To my knowledge, such large holes that are not covered by the density of the HK97 fold have not been reported in other tailed phages.

Coloring the density that is 2 Å away from the fitted gp5 coordinates shows large density that is not formed by the HK97 fold of the MCP (Figure 22C & D and Figure 23C & D). The outer surface of the colored map shows three types of extra density which fit exactly to the surface features previously discussed: the dimeric density near the local 2-fold symmetry of two neighboring capsomers highlighted in red; the knob-shaped density at the center of hexamers highlighted in yellow; and the flexible density at the global 2-fold symmetry highlighted in dark

green. The inner surface only shows one type of extra density, a pentameric structure sitting under the pentamer colored in green.

An important question to ask is whether the four types of extra density are formed by the roughly 100 N-terminal residues of gp585* that is not included in the predicted HK97 fold, or by decoration proteins that strengthen the capsid structure but are usually dispensable in laboratory conditions. Apparently the flexible density attached to the global 2-fold symmetry and the pentameric density attached to the pentamer cannot be formed by gp585, because they only display on specific regions and are simply too large to be formed by the N-terminal. The extra density at the center of the hexamer is very likely to be formed by a decoration protein, because gp585 may have a similar 3D structure to gp24 and the T4 hexamer has the hoc density at the same position with a similar shape. The dimeric density is at least not formed only by the N-terminal of gp585. A rough estimate of the volume of each monomer is a little more than half of the volume of the HK97fold in each subunit, much larger than a structure only formed by the N-terminal of gp585. However, because the 2-D structure prediction shows matching between the N-terminal of gp585 and the insertion domain in gp24 and the insertion domain is located near the local 2-fold symmetry, the N-terminal of gp585 may form a similar structure and interact with the decoration protein.

The positions of the four types of decoration density may provide insights into their roles in the formation of the capsid. The dimeric density and the knob-shaped density together cover the holes left by the density formed by the HK97 fold, so they may play important roles in forming the whole capsid shell. The dimeric density may strengthen the inter-capsomeric interaction because the monomer slightly extends to the neighboring capsomer. The pentameric density may be involved in stabilizing the pentamer because each monomer of this density

contacts two neighboring subunits in the pentamer. The flexible density at the global 2-fold symmetry is likely the result of icosahedral averaging of the proximal part of the hair around the capsid.

Figure 22 Local views of the density of HK97 fold and the decoration density

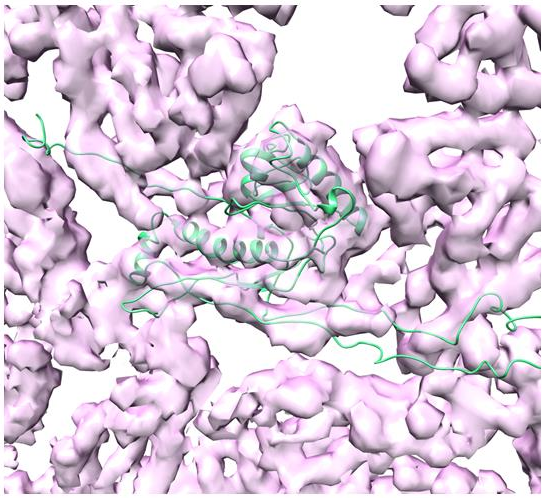
A. The HK97 gp5 structure fitted into a subunit of a hexamer in the 121Q cryo-EM model. Only the density near the fitted coordinates of gp5 is shown. The density far away from the coordinates is hidden.

B. Density of the asymmetric unit around the fitted gp5 coordinates. An asymmetric unit of the T=28 capsid consisting of 28 subunits, including one subunit from the pentamer and four and a half hexamers is colored in pink. The fitted gp5 coordinates are not displayed. The density around the gp5 forms the majority of the capsid shell but leaves two kinds of holes indicated by arrows, one in the middle of the capsomer and one at the local 2-fold symmetry between two neighboring capsomers.

C. Decoration density on the outer surface of the same shell density in B. The extra density far away from the fitted gp5 coordinates are colored in red or yellow according the position of the density. The extra density at the global 2-fold symmetry is not displayed, since we think this is the attachment site for head fibers.

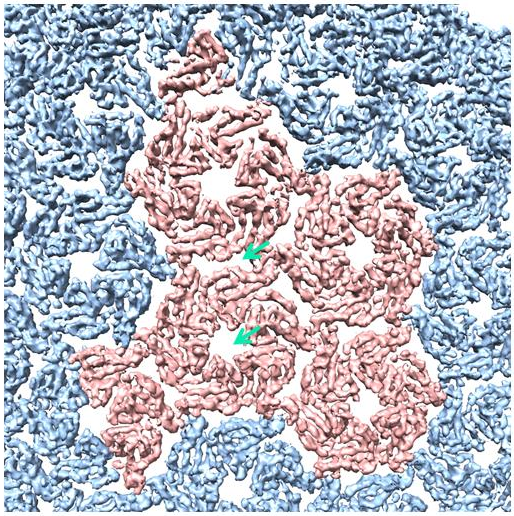
D. Decoration density on the inner surface. The only decoration density observed is the 5-mer structure under the pentamer. The red and yellow densities are on the outside (as in C), but visible because of the holes.

A



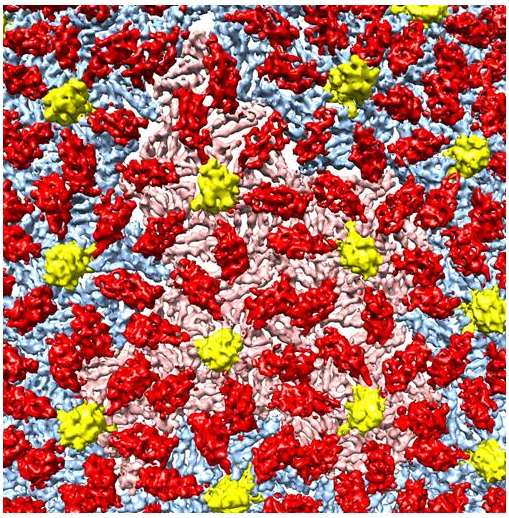
Fitting gp5 structure in the density of a 121Q subunit

B



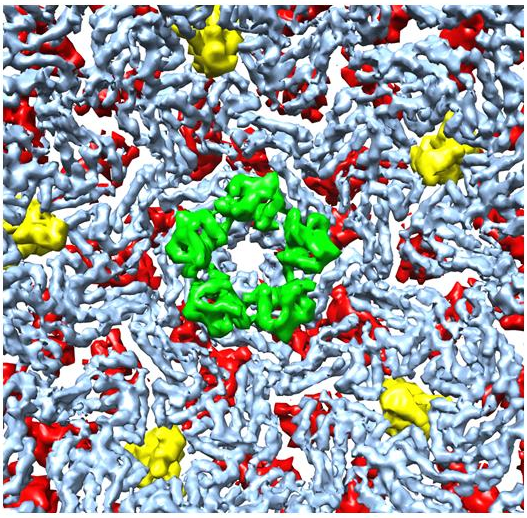
Density of the HK97 fold of an asymmetric unit

C



Decoration density on the outer surface

D



Decoration density on the inner surface

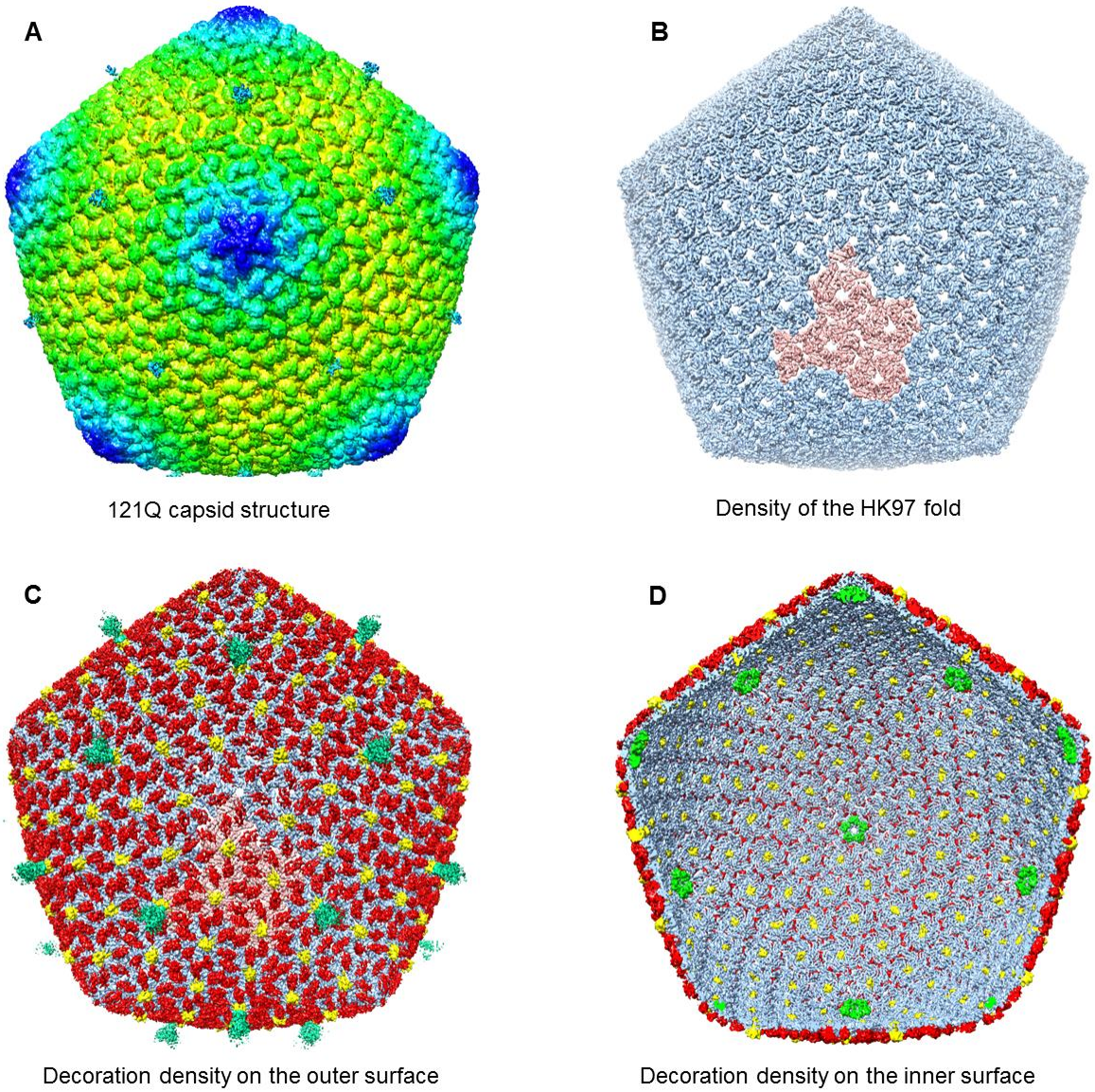
Figure 23 Density of the HK97 fold and the decoration density

A. Surface of 121Q capsid color by radii.

B. Density around the fitted gp5 coordinates.

C. Decoration density on the outer surface. The coloring is the similar to Figure 22C except that the flexible at the global 2-fold symmetry is displayed here.

D. Decoration density on the inner surface.



3.3.4 Structure of the inner body and the tail

The inner body, first found in the T=27 phage phiKZ (Wu, Thomas et al. 2012), is a cylindrical protein core inside the DNA-filled capsid of some large phages. The presence of the inner body can be detected by applying a high dosage of electron beam which the inner body is more sensitive to than the dsDNA around it. The 121Q capsid exposed to heavy electron beam displays "a bubble-gram" of the initial beam damage inside the capsid (Figure 24A). The structure damaged by the electron beam is about 132 nm long and roughly around the portal axis. The size, shape and position of the beam damage in 121Q are very similar to what was observed in phiKZ, so the 121Q capsid may contain an inner body with similar structure. The inner body does not have icosahedral symmetry, so the structural information about the inner body is lost in the previous capsid map reconstructed with icosahedral averaging. The asymmetric reconstruction of the 121Q capsid shows a 20 nm X 100 nm structure inside the capsid. A section of the density map (Figure 24B, upper) shows the density of the inner body which is as dense as the capsid shell at the center but does not have a well define boundary with the DNA. A section of the map with very high contour level gives a rough outline of the inner body (Figure 24B, lower). However, this structure is probably not an accurate description of the inner body. The technical difficulty in precisely locating the inner body in the DNA-filled capsid results in data averaging along the portal axis which is not perfectly aligned with the inner body.

The tail structure is resolved with both helical symmetry and 6-fold rotational symmetry. The former reconstruction gives better resolution but loses the structure that does not have helical symmetry. The length of the tail including the neck is 118 nm, the shortest among our jumbo phage collection. The width of the sheath is about 27.6 nm. The repeat distance between two nearby sheath rings is about 4.1 nm. Some extra flexible density can be observed at the two

ends and in the middle of the tail part covered with sheath protein. This flexible density may be the averaged density of the proximal end of the hair attached to the tail.

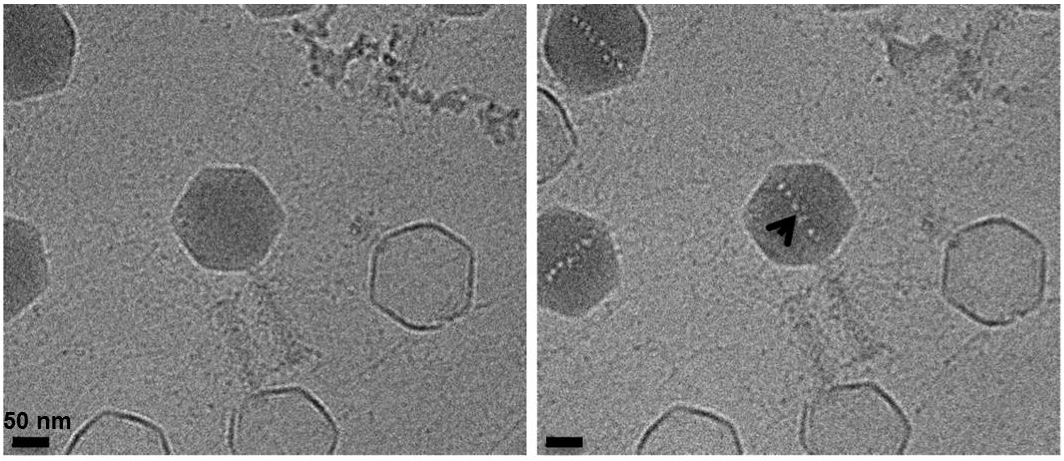
Figure 24 Structure of the inner body and the tail

A. Cryo-EM images of the same area of the frozen grid of 121Q phages before or after a high dose of electron beam. The protein core inside the capsid is the most susceptible to beam damage, which can be used to detect the presence of the inner body.

B. Structure of the inner body displayed in the asymmetric reconstruction of 121Q capsid. In the upper figure, the section of the density map shows densely packed protein core forming a long darker area between the portal and the distal vertex. In the lower figure, a slice of the map with high contour level outlines the structure of the inner body.

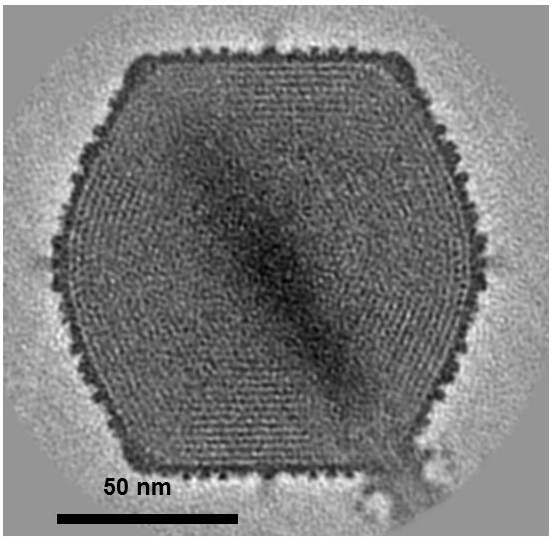
C. Tail structure. The tail reconstruction using helical symmetry is shown on the left. The tail reconstruction using C6 symmetry is shown on the right. The helical reconstruction has a high resolution, but lost the density such as flexible density around the C6 model. The capsid structure shown here has a low resolution because it is reconstructed together with the tail based on the tail signal. The scale in this figure is the same as the scale in B.

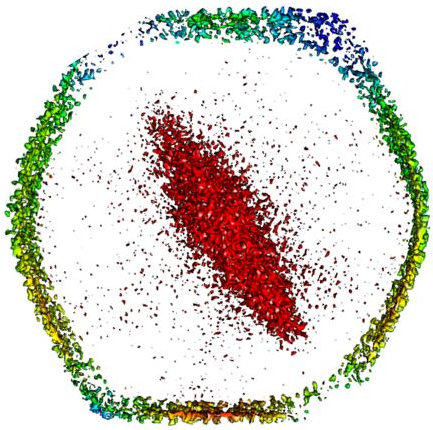
A



The some area of the cryo grid of 121Q phages before or after heavy beam exposure

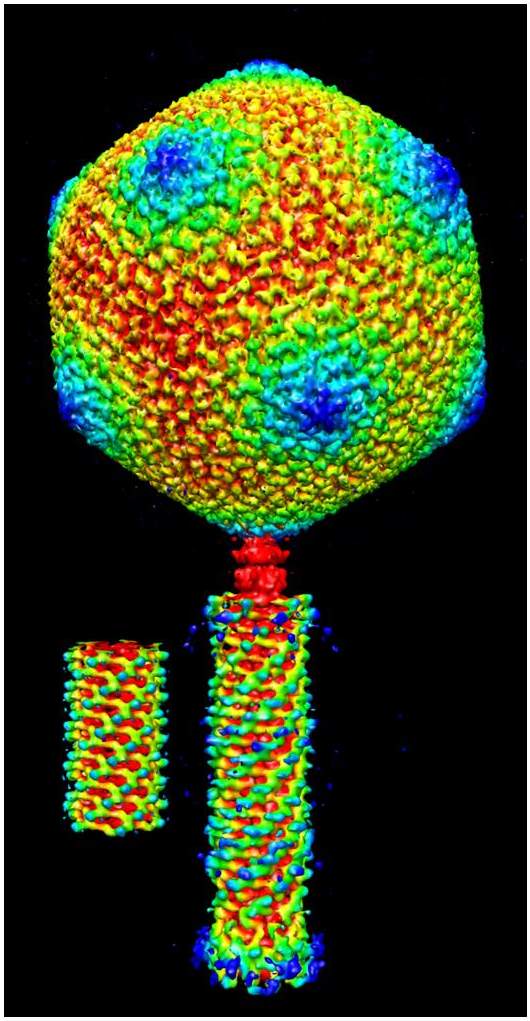
B





Inner body in the asymmetric model

C



Tail structure with helical (left) or rotation (right) symmetry applied

3.4 STRUCTURAL PROTEINS OF 121Q

3.4.1 Identification of decoration proteins

Attempts were made to identify the decoration proteins. Because empty capsids cannot be completely separated from cell debris and tails, most structural proteins of 121Q cannot be identified as capsid or tail proteins by comparing the band intensity of a SDS gel with capsid and tail samples loaded. Since urea can successfully extract decoration proteins and internal proteins from the empty capsid of PAU, the empty capsid of 121Q was also treated with several conditions hoping that some proteins would be extracted or digested by the applied condition. The empty capsid sample was mixed with the reagent and incubated at room temperature or 37 ℃ for a few hours. Then the sample was diluted and the capsid concentrated by ultracentrifugation was prepared for SDS PAGE. Proteins that came off the capsid during the treatment would stay in the supernatant and show reduced band intensity in the gel. Unfortunately, loss of protein is not observed when the lane of the treated capsid is compared to the control (Figure 25A). The gel shows that no protein was extracted by urea; trypsin was not able to digest any flexible protein; and the amount of the host protein seemed not affected by either Triton X-100 or EDTA which was a test for weakening interaction between the capsid and the cell debris.

Three proteins, gp599, 600 and 552, can still be identified as capsid proteins based on the gel and the mass spectrometry data of the total capsid protein. On the gel, three heavy bands in addition to the MCP band can be easily recognized as capsid protein bands: one is about 150 kDa and the two are about 20 kDa based on the marker (Figure 19B). The copy numbers per capsid for these proteins are about 26 copies per capsid for the large protein and 415 and 764 for the

two smaller proteins, estimated by comparing the band intensity to the intensity of the MCP of which each capsid has 1675 copies. From the mass spectrometry data, three proteins, gp599, 600 and 552, have comparable numbers of total spectrum counts compared with the MCP, much more than the rest of the structural proteins in the empty capsid sample. The size of these three proteins matches the size of the three heavy bands. Therefore, gp599 and gp600 are probably the two small proteins and gp552 is probably the large protein.

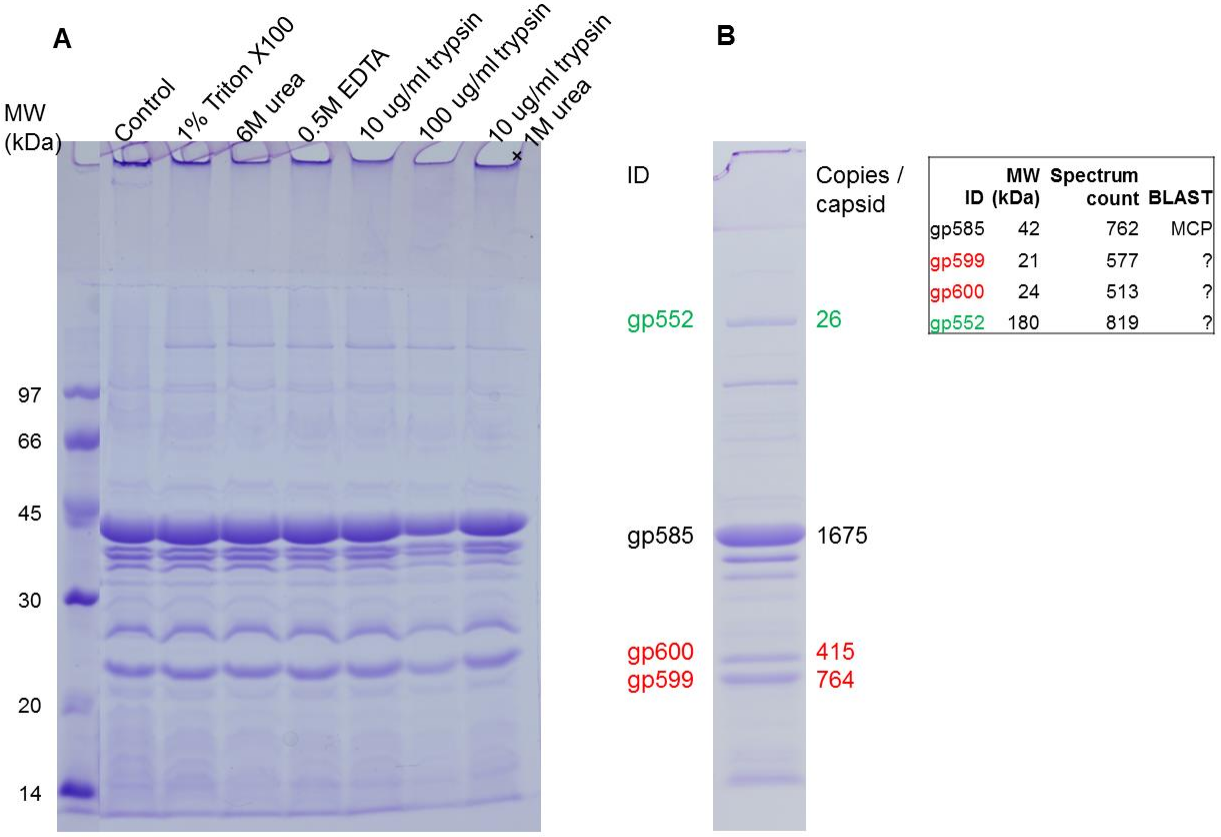
The size and copy number per capsid of gp552 suggest that gp552 is a good candidate for the hair protein. If there is a hair attached to every global 2-fold symmetry on the 121Q capsid and every hair contains only a single protein, the hair protein is a very large protein and each capsid should contains 30 copies of it. However, the hair might consist of more than one peptide and not all global 2-fold symmetry positions are occupied by the hair.

The size and copy number of gp599 and gp600 suggest that gp599 and gp600 may form the dimeric decoration structure at the local 2-fold symmetry. The estimated volume of each monomer is about half the volume of a MCP subunit and each capsid has the same number of copies of monomer as the MCP. If the monomer is formed by a single protein, each capsid will have 1675 copies of an approximately 20 kDa protein. Because both gp599 and gp600 are about 20 kDa, they have sequence homology (31% identity) and each capsid has about 1177 copies of the two, the dimeric decoration structure may be formed by a gp599 peptide and a gp600 peptide, possibly in the form of a heterodimer.

Figure 25 Identification of decoration protein

A. SDS gel of 121Q empty capsids treated with reagents. Some bands have strange shapes due to the problem of the acrylamide used in the gel. For unknown reasons, a few high molecular weight bands are missing in the control lane.

B. Identification of three capsid proteins. On the left shows the capsid sample lane in (Figure 19B) with the number of copies per capsid. On the right are the protein candidates identified by mass spectrometry according to the abundance and size. Gp45 has the fifth largest spectrum count in the whole sample mass spectrometry experiment.



SDS gel of treated capsids

Gel quantification and mass spectrometry of capsid protein

3.4.2 Capsid proteins and tail proteins

Another goal in the 121Q project is to identify the large number of capsid proteins and tail proteins in the 121Q virion. Because the purification of the empty capsid sample and the tail sample was not so successful and the resolution of the SDS gel was low, identifying proteins based on gel bands is inaccurate and inefficient. Instead, all the proteins in the sample were identified in a single mass spectrometry experiment and capsid or tail proteins were identified by comparing the abundance of the protein in the capsid sample and in the tail sample. The normalized spectrum count of a protein was used to quantify the relative abundance of the protein in the sample. The ratio of the normalized spectrum count of a protein from the capsid sample to the count from the tail sample (or C/T ratio) was used as the criterion to determine the location of the protein. This method may not be accurate, but it is the only tool within our reach to provide a general idea of how many of the putative proteins in the 121Q are capsid or tail proteins and how these structure proteins are distributed in the genome. 121Q proteins are listed in descending order of C/T ratio in Table 3.

The threshold for calling a capsid or tail protein was based on the spectrum count ratio of the structural proteins identified by biochemistry or by protein BLAST (Table 3). The ratios for the identified capsid proteins (protease, portal, MCP and the two dimeric decoration proteins) range from 2.40 to 5.96, so proteins with spectrum count ratios equal or larger than 2.40 are considered capsid proteins. Similarly, proteins with ratios equal or smaller than 0.42, the multiplicative inverse of 2.40, are consider tails proteins. This threshold for tail proteins is consistent with the ratios of several tail proteins identified by BLAST (the tail sheath stabilizer, baseplate wedge, sheath, baseplate hub and two tail fibers). Out of 83 structural proteins, 26 are considered as capsid proteins according to the above criteria, 32 structural proteins are

considered tail proteins and 25 proteins with the ratio between 2.39 and 0.41 are undecided. However, the 180 kDa protein gp552 is an exception. The SDS gel (Figure 19B) and mass spectrometry data suggests that both the capsid and the tail contain a large amount of gp552 which is identified as the hair protein on the capsid. So the unique hair structures around the capsid and the tail of 121Q may be formed by the same protein.

The structural protein encoding genes are not randomly distributed in the genome but generally grouped together. The 121Q genome map with structural genes marked in different colors shows that most genes that encode capsid proteins are in a 56 kb region between 331 kb and 38 kb in the circularly permuted genome. Most genes that encode tail proteins are located in a 99 kb region between 232 kb and 331 kb (Figure 26). Most of the genes in the middle of the map do not encode structural proteins present in the mature phage. Most of these genes only share homology to genes in a closely related group of four phages encoding hypothetical proteins. A few of them encode proteins involved in DNA metabolism. However, the region of 118 – 131 kb contains genes encoding eleven structural proteins that have similar sequence and size. The role of these similar structural proteins in the 121Q phage is not clear.

The classic gene cluster with consecutive genes encoding the portal, protease, scaffolding protein and MCP is conserved in 121Q. Genes encoding portal, protease and MCP can be identified by BLAST as ORFs 580, 583 and 585. The ORF584 is very likely to be the scaffolding gene because the first 260 residues of the 371 AAs gp584 have a very high percentage of predicted coiled coils (data not shown). Two additional genes are inserted between the genes encoding the portal and the protease, both encoding structural proteins. Therefore, the classic gene cluster is more preserved in 121Q compared to the T=25 phage PAU.

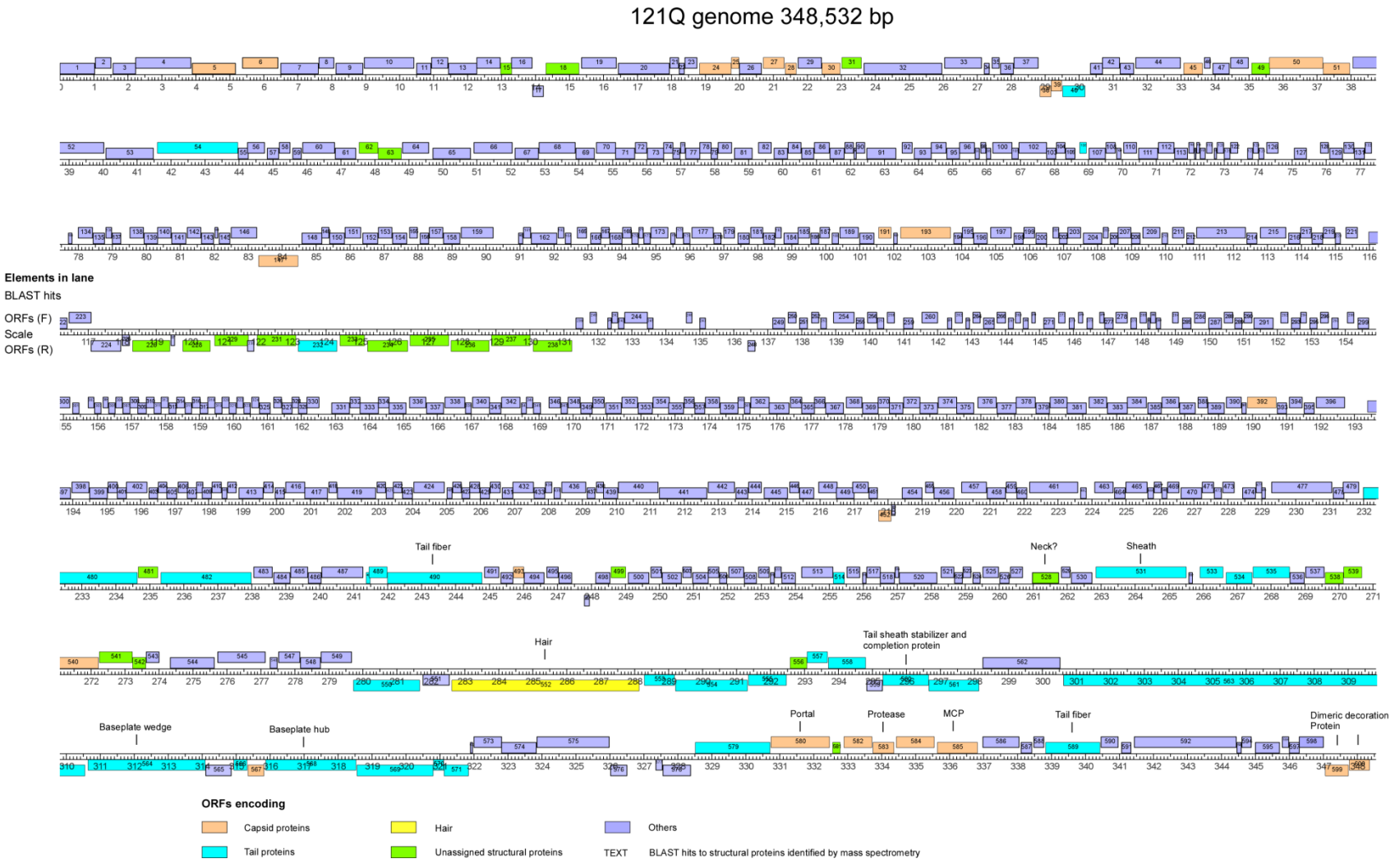
Table 3 List of structural proteins identified mass spectrometry.

The list is sorted in descending order by the ratio of the spectrum count ratio (C/T ratio). The proteins are listed in three separated columns, the predicted capsid proteins on the left (C/T ratio $\geq$ 2.4), the undecided structural proteins (0.42 < C/T ratio < 2.4) and the predicted tail proteins (C/T ratio $\leq$ 0.42). Five proteins with less than three spectrum count in total are ignored. Gp552 and gp563 formed lagging bands after a brief electrophoresis into the lower SDS gel. The detailed mass spectrometry data are listed in (**Table 7**).

ID	MW (kDa)	C/T ratio	Protein	ID	MW (kDa)	C/T ratio	Protein	ID	MW (kDa)	C/T ratio	Protein
5	48	∞		49	21	2.34		566	15	0.42	
6	39	∞		499	17	2.27		40	25	0.41	
24	35	∞		538	21	2.24		232	41	0.38	
25	8	∞		63	26	2.16		561	57	0.38	
30	21	∞		541	36	2.16		560	52	0.36	Tail sheath stabilizer & completion
38	13	∞		542	14	2.13		557	22	0.35	
493	12	∞		528	30	1.89	Neck?	533	25	0.35	
567	19	∞		62	21	1.81		534	29	0.33	
584	42	∞		233	24	1.65		555	41	0.32	
45	21	6.93		539	20	1.54		553	31	0.31	
51	29	6.38		237	41	1.27		558	40	0.30	
583	23	5.96	Protease	15	13	1.22		535	40	0.30	
50	60	3.88		235	41	1.10		554	78	0.29	
452	14	3.11		18	37	0.96		482	96	0.29	
147	43	3.08		226	39	0.85		54	90	0.29	
600	21	3.07	Dimeric surface protein	229	35	0.83		564	131	0.28	Baseplate wedge
193	55	3.00		234	43	0.79		579	80	0.26	
540	57	3.00		581	9	0.78		480	99	0.25	
582	31	2.96		236	39	0.72		531	97	0.23	Sheath
599	24	2.83	Dimeric surface protein	238	41	0.71		570	12	0.22	
191	15	2.76		228	30	0.63		569	85	0.22	
580	65	2.56	Portal	556	19	0.59		550	70	0.21	
27	23	2.52		31	21	0.58		488	5	0.21	
39	12	2.45		481	21	0.53		489	19	0.19	
28	12	2.43		231	40	0.44		568	96	0.19	Baseplate hub
585	42	2.40	MCP					589	57	0.15	Tail fiber
								552	180	0.14	Hair
								563	382	0.08	Tail fiber?
								571	28	0.07	
								490	105	0.04	Tail fiber
								106	8	0.00	
								514	12	0.00	

Figure 26 Map of the 121Q genome

The map is generated by DNA Master by Dr. Jeffrey Lawrence and edited using Canvas. ORFs encoding structural proteins detected by mass spectrometry are displayed in different colors based on the predicted location of the protein, indicated on the bottom. Homology of the structural genes identified by BLAST is labeled in the map. Only BLAST matches to genes encoding the structural protein detected by mass spectrometry are shown in the map.

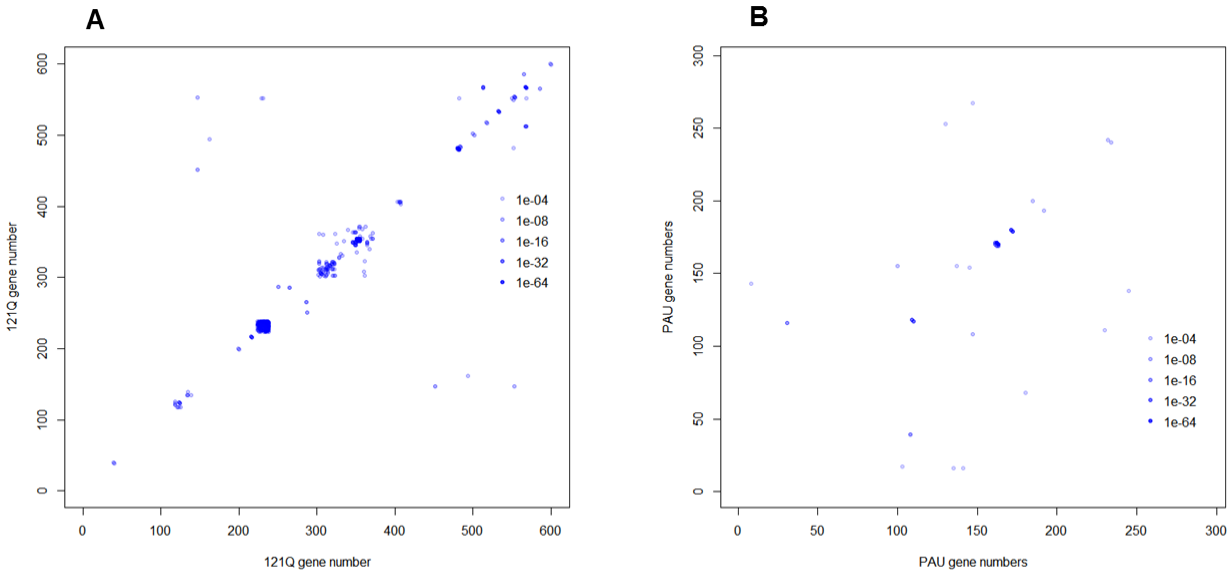

121Q genome 348,532 bp
Elements in lane
BLAST hits
ORFs (F)
Scale
ORFs (R)
Tail fiber
Neck?
Sheath
Hair
Tail sheath stabilizer and completion protein
Baseplate wedge
Baseplate hub
Portal
Protease
MCP
Tail fiber
Dimeric decoration Protein
ORFs encoding
Capsid proteins
Hair
Others
Tail proteins
Unassigned structural proteins
TEXT
BLAST hits to structural proteins identified by mass spectrometry

Lots of the putative genes in the middle of the genome are short and similar in size to the nearby genes. My guess for such a repeated pattern of non-conservation genes is that the 121Q capsid has been expanding the length of its genome by gene replication, which is supported by the homology between 121Q genes measured by BLASTing each peptide sequence against the 121Q peptide library. If a group of genes are duplicated from a single gene, these genes may have sequence homology in addition to the similar length. Homology between two genes is shown in a 2-D matrix with both X and Y axis representing the gene ID. If sequence homology (set to a threshold of E-value $\leq 10^{-4}$) to the query peptide is found in a few candidates, a marker at the coordinate of the ID of the query peptide and the ID of the candidate with the smallest E-value is drawn in the 2-D matrix. If a group of nearby genes have sequence similarity, multiple marker will be displayed near the diagonal. Figure 27A shows three groups of genes with homology, including genes 114 – 132, 223 – 238 and 308 – 380. Much less frequent makers are observed in the region where the structural gene is located and in the PAU genome as a control.

Figure 27 Gene duplication in 121Q genome

A. 2-D matrices of homologous pairs in 121Q genome

B. 2-D matrices of homologous pairs in PAU genome as a control



Homologous pairs in the 121Q genome

Homologous pairs in the PAU genome

3.5 DISCUSSION

3.5.1 A jumbo phage similar to other phages

Compared to PAU, 121Q is a jumbo phage more evolutionarily connected to other phages. Its connection to other well-studied phage is first reflected by the genome sequence. A handful of 121Q genes show sequence homology to conserved genes in well-studied middle-sized phages. The conserved capsid gene cluster is nearly intact, with only two structural genes inserted between the portal and the protease. The MCP shows similarity in primary sequence to a number of major capsid proteins. The predicted 2D/3D structure shows great similarity to T4 gp24, not only in the conserved region that forms the core HK97 fold, but also in the N-terminal region that forms the peripheral domains including the insertion domain and the small helix-turn-helix domain.

121Q is also evolutionarily connected to a few phages that were recently isolated and sequenced. These phages include but are not limited to:

1. PBECO4 (Kim, Hong et al. 2013), isolated from sewage in Korea, infecting Escherichia, 348 kb;
2. vB_KleM-RaK2 (Simoliunas, Kaliniene et al. 2013), from sewage in Lithuania, infecting Klebsiella, 346 kb; and
3. vB_CsaM_GAP32 (Abbasifar, Griffiths et al. 2014), from sewage in Lithuania, infecting Cronobacter, 359 kb.

The genomes of the four phages have similar lengths and really good sequence homology to 121Q. Matches are found for most genes in the 250 – 50 kb region of the 121Q genome where

most of the structural genes are located. Therefore, 121Q probably belongs to a group of closely related T=28 phages living in sewage worldwide.

3.5.2 A "hairy" phage

It is not unusual to observe fibers located on places other than the far end of the tail that infects the host. For example, T4 has the 40 nm whisker that extends outward from the collar end of the tail. The whisker is not essential but can prevent adsorption of the tail fiber to host cells under unfavorable conditions (Conley and Wood 1975). Phage ϕ29 has the 11 nm short head fibers attached to the local 3-fold symmetry of the capsid. The head fiber is dispensable in laboratory conditions (Salas, Vasquez et al. 1972).

The 121Q hair is probably a new type of non-tail fiber. The length of the hair, approximately 100 nm, is comparable to long tail fibers. The hair is attached at some unusual locations for phage fibers, not only on global 2-fold symmetry axes of the capsid, but also on the middle and the two ends of the tail section covered with sheath. Both the tail hair and the capsid hair are probably formed by the same protein gp552. It is interesting that while gp552 binds to both capsomers and sheath it also only binds to a specific type of hexamer that only has subtle conformational differences to the other hexamers.

121Q phages stored in the cold slowly lose hairs suggesting that the hair is probably dispensable. Gp552 is not quite conserved in the four sewage jumbo phages, and homology is not detected to tail fibers which are considered conserved. The difficulty in removing cell debris from the 121Q particles suggests that the hair may be interacting with the host cell wall during infection, facilitating the phage adsorption processes. This feature may provide some

evolutionary advantages to a sewage phage because it helps the phage quickly target the host instead of being rushed away by the running water.

The tail also contains a huge 360 kDa protein gp563 which is conserved in the four sewage jumbo phages. The number of gp563 molecules per tail is about 1/6 of gp552. The role of gp563 is not clear.

3.5.3 Essential minor capsid proteins?

We describe the additional proteins of the capsid shell of 121Q as "decoration proteins" which implies that they are nonessential, instead of "minor capsid proteins" which emphasizes that they are less abundant than the MCP. The data, however, suggest that some of the decoration proteins may be essential. We suggest that the dimeric structure at the local 2-fold symmetry axes and the knob-shaped structure at the middle of the hexamer may be essential because without them, the capsid shell could have holes big enough to allow DNA release. Further, the interactions between the decoration proteins and the capsid shell may be quite strong, since 6 M urea could not extract either gp599 or gp600 from the empty capsid.

Essential minor capsid proteins are found in other viruses, including P30 in the enveloped phage PRD1 (Rydman, Bamford et al. 2001) and several minor proteins in the human herpes simplex virus (Toropova, Huffman et al. 2011), but not in tailed phages. Also we cannot find any other examples in tailed phages where the MCP (and the vertex protein) does not form a continuous and intact capsid shell. An essential role for the 121Q decoration proteins in the formation of a tailed phage would be a novel discovery, but proof of this hypothesis will require further experiments.

3.5.4 Gene duplication in the genome

Jumbo phages generally have low DNA packing density and high terminal redundancy. However, compared to the T=25 phage PAU and the T=27 phage PBS1, the DNA packing density of 121Q is relatively higher and the terminal redundancy of 121Q is much lower (Table 4). This increase in genes/unit volume suggests that the extra genetic diversity of 121Q may confer some kind of selective advantage. The genes that expand the size of the 121Q genome are likely those genes predominantly of unknown function in the middle of the genome map. Most of these genes are short in length and transcribed in the forward direction, except the 118 kb - 131 kb region which contains relatively larger genes encoding structural proteins in the reverse direction. The shared sequence homology and similar length of a few groups of adjacent genes suggests 121Q has expanded its genome length through gene duplication. The degree of gene duplication may be underestimated in Figure 27 because E-value as the criterion for homology search favors long sequence but most genes in the middle of the genome are short.

Gene duplication is an important mechanism of creating genetic novelty in evolution. Genetic redundancy generated by gene duplication is usually not restricted by selective pressure, providing fuel for genetic innovation (Ohno 1970). Alternatively, all copies of the duplicate genes may freely acquire mutations as long as all detrimental mutations are complemented (Force, Lynch et al. 1999). This process may provide adaptive benefits because it allows partition of the original gene and specialization of the subfunctions of the gene. Gene duplication may be a way for 121Q to take advantage of the large capsid to accumulate beneficial effects.

4.0 CHARACTERISTICS OF DNA IN JUMBO PHAGES

4.1 INTRODUCTION

The capsid of the tailed phage consists of two major parts: the highly compressed chromosomal DNA and the capsid shell that serves as the DNA container. Choosing a properly sized chromosome is far more critical to a phage than to a cellular organism. To survive, the phage genome should include all the essential genes and other genes that improve fitness, but its capsid imposes a physical limit to the amount of DNA that the capsid is capable of carrying. A larger capsid allows a bigger genome potentially with more beneficial genes, but at the cost of more building materials for each virion and a reduced burst size. In mid-sized and jumbo phages which use the headful DNA packaging strategy, the circularly permuted chromosome DNA needs to maintain redundant termini to avoid losing important genes during phage replication. Therefore, the chromosome size is a result of coordination between the genome, the capsid and the DNA packaging.

The successful application of biophysical techniques in phage research in the past decades, particularly in cryo-EM imaging and reconstruction, has greatly expanded our knowledge of the protein capsid shell in various phages. In the meantime, progress in understanding how phages manage the chromosomal DNA inside the capsid is less abundant. The jumbo phage collection in our lab provides a nice opportunity to learn how tailed phages

manage DNA in a very large capsid compared with small and mid-size phages. In the paper about the isolation of the *Sphingomonas paucimobilis* PAU, Achermann et al reported the DNA size of PAU to be 412 kb, determined by pulsed-field gel electrophoresis (PFGE) (Ackermann, Auclair et al. 1994). We found that this number was incredibly high in light of the genome size determined by sequencing. If it were true, the terminal redundancy of PAU DNA would be over 80% and the DNA packing density would be much higher than that in phage λ. However, we did predict that the DNA of PAU had a sizable terminal redundancy, because the genome size of the T=25 phages PAU (219 Kb) is close to the genome size of the smaller T=19 phage N3 (207 Kb) but significantly smaller than the slightly larger T=27 phage PBS1 (252 Kb). Investigating the properties of PAU DNA was the initial motivation of my jumbo phage DNA project, because such an exceptionally large terminal redundancy, if it is true, might be a result of a recent evolutionary event that increased capsid size.

This chapter covers my work on the properties of jumbo phage DNA. The DNA modification of PAU DNA was analyzed by restriction digestion and nucleoside HPLC. The length of jumbo phage DNA was determined by two-dimensional pulsed-field gel electrophoresis. The chromosome size, terminal redundancy and DNA packing density in phages with a variety of T-numbers are compared and discussed. A negative correlation of DNA packing density and capsid size is discovered and a simple model is proposed to explain this observation.

4.2 DNA MODIFICATION IN JUMBO PHAGES

A likely explanation for the unusually high PFGE measurement is that the PAU DNA may contain base modification that greatly slows the DNA migration during electrophoresis. Because the presence of modified bases in the site of a restriction enzyme may stop or slow down its activity, we used five restriction enzymes to cut PAU DNA overnight and compared the gel pattern of the digestion with the theoretical gel pattern after complete digestion (Figure 28A & B). BstUI (CGCG), HhaI (GCGC) and SspI (AATATT) have only C/G or A/T in their recognition sites and are susceptible to cytosine or thymine modification in phage DNA (Huang, Farnet et al. 1982). TaqI (TCGA) and EcoRV (GATATC) have all four nucleotides in their recognition sites, but TaqI can extensively cut modified phage DNA (Huang, Farnet et al. 1982). The agarose gel shows that BstUI and HhaI could not cut PAU DNA at all. SspI cut the PAU DNA extensively and recognizable bands were formed, but the sizes of the bands are higher than the expected size in the simulation. TaqI slowly cut the phage DNA, resulting in a wide smearing region in the lane. For unknown reason, most of the phage DNA incubated with EcoRV was trapped in the well. A smearing faint band at a similar position of the DNA treated with BstUI and HhaI suggests that EcoRV may not be able to cut PAU DNA. The restriction digestion experiment suggests that PAU DNA contains modification in cytosine or guanine, which slows down the migration of the DNA fragments in the agarose gel.

I used a reversed-phase HPLC method for identifying and purifying the modified nucleoside. The DNA produced by phenol extraction of CsCl-purified phages was sequentially digested by DNase I, Nuclease P1 and bacterial alkaline phosphatase. The HPLC separation of the major nucleosides were performed on a single C18 column with an acidic $NH_4H_2PO_4$ buffer pair, which was adapted from (Kuo, McCune et al. 1980). Nucleosides in the mobile phase were

monitored by UV absorption in the range of 220 ~ 320 nm. The chromatography of the plasmid DNA prepared by mini-prep showed four peaks with maximum absorption around 265 nm. These peaks were identified as dC at 8 min, dG at 16 min, dT at 23 min and dA at 40 min, based on their relative positions and spectra (Figure 28C, top). The chromatography of the PAU DNA showed an extra peak at 46 min in addition to the four peaks of the unmodified nucleosides (Figure 28C, bottom). The spectrum of the extra nucleoside is similar to that of dC, with the wave length of the maximum absorption at around 270 nm and a high absorption at the far UV end (Figure 28D). The chromatography of PAU DNA also displays a significantly smaller dC peak compared with the the dG peak, contrary to the common case that the double-stranded DNA has equal amount of cytosine and guanine. However, the dC peak and the extra peak together can roughly match the dG peak. Therefore, the extra nucleoside is probably a modified dC.

The analysis based on the chromatography and restriction digestion data suggests that the cytosine in PAU DNA is partially modified. The modifying group may be large and hydrophobic, because the dC* peak is at the right end in the chromatography which requires high affinity to the C18 substance. To learn the structure of the modification, we collected the dC* peak during the HPLC experiment and tried to concentrate it for mass spectrometry. The modified cytidine was an unstable molecule in my hand. Its concentration measured by HPLC declined rapidly in all storage conditions we tested. It was completely unrecoverable after being dried out, regardless of the choice of buffer or container. A concentrated sample was prepared using nitrogen flow which might stabilize the modified cytidine, and sent to the mass spec facility in Department of Chemistry. However, the result we received weeks later was very noisy and hard to interpret.

The HPLC analysis of the nucleosides of phage G DNA shows that the cytosine in G DNA was also partially modified, but the percentage of the modified cytosine is less than 50%.

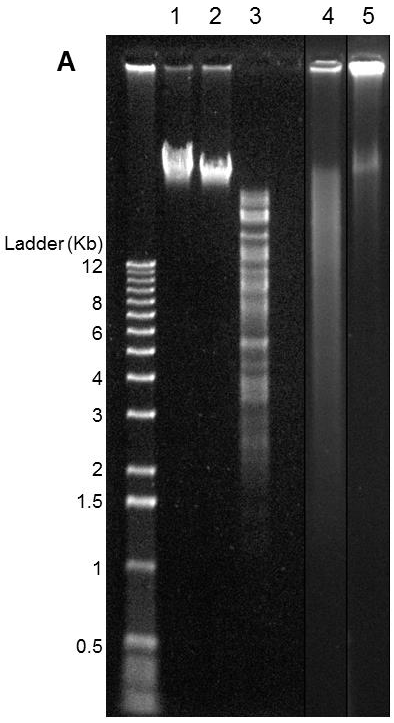
Figure 28 Restriction digestion and nucleotide analysis of PAU DNA

A. Restriction digestion of PAU DNA. The sites of the restriction enzymes are listed on the right side of the figure. Lanes 4 and 5 are moved from another gel with scaling to fit the current gel. The ladder is the Invitrogen 1 Kb ladder.

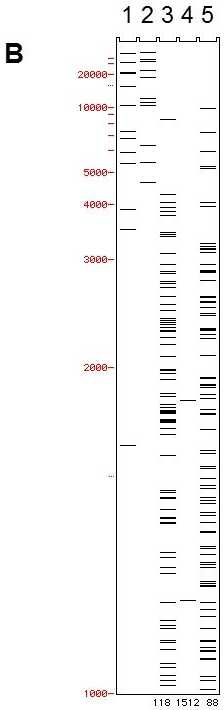
B. Simulation of restriction digestion of PAU DNA using the same restriction enzymes. A bar represents a piece of DNA fragment after the complete restriction digestion. The number at the bottom of each lane indicates the number of omitted DNA fragments that are smaller than 1 Kb.

C. Separation of nucleosides in plasmid DNA (top) and PAU DNA (bottom) by C18 chromatography. The plasmid DNA and phage DNA were hydrolyzed with DNase I, Nuclease P1 and bacterial alkaline phosphatase. For both samples, the time curve of 265 nm absorption and the UV spectra of the major peaks found from the time curve are shown. The scales of the spectra are not uniform. The two peaks before 8 min with weak absorption at 265 nm are junk. The peaks were identified by their position and spectra. The PAU DNA data show an extra nucleotide at 46 min and a weaker dC peak compared with the dG peak.

D. UV spectra of the five nucleotides peaks from digested PAU DNA. The figure shows an imbalance between the dC and dG peaks, and a similarity of the wave length of maximum absorption between dC and the extra nucleoside. So the extra nucleoside may contain a modified form of cytosine.

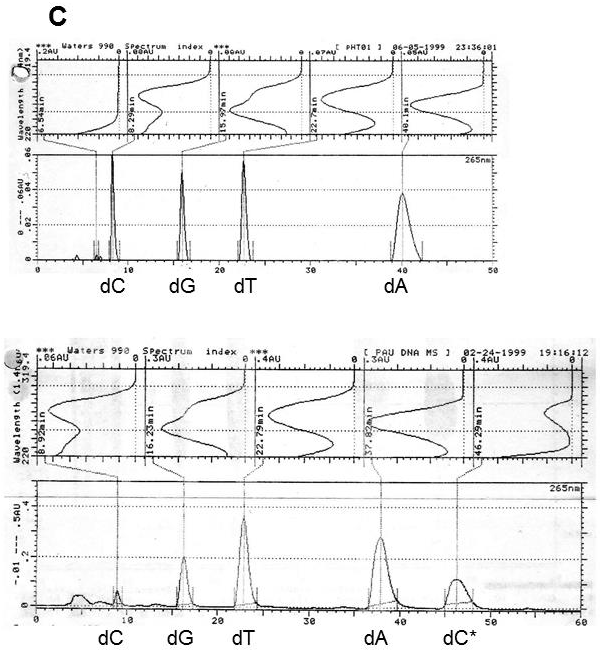


Restriction digestion of PAU DNA

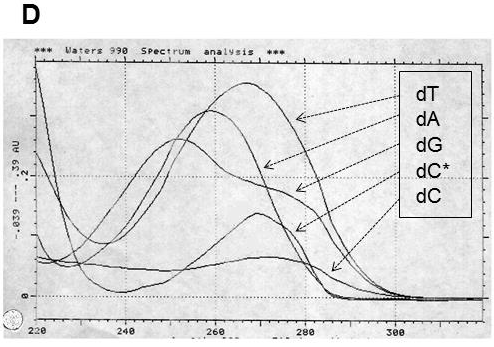


Enzymes	Sites
1. BstUI	CGCG
2. HhaI	GCGC
3. SspI	AATATT
4. TaqI	TCGA
5. EcoRV	GATATC

Simulation of restriction digestion of PAU DNA



C18 HPLC data of nucleosides in plasmid DNA (top) and PAU DNA (bottom)



Spectra of the peaks from PAU DNA

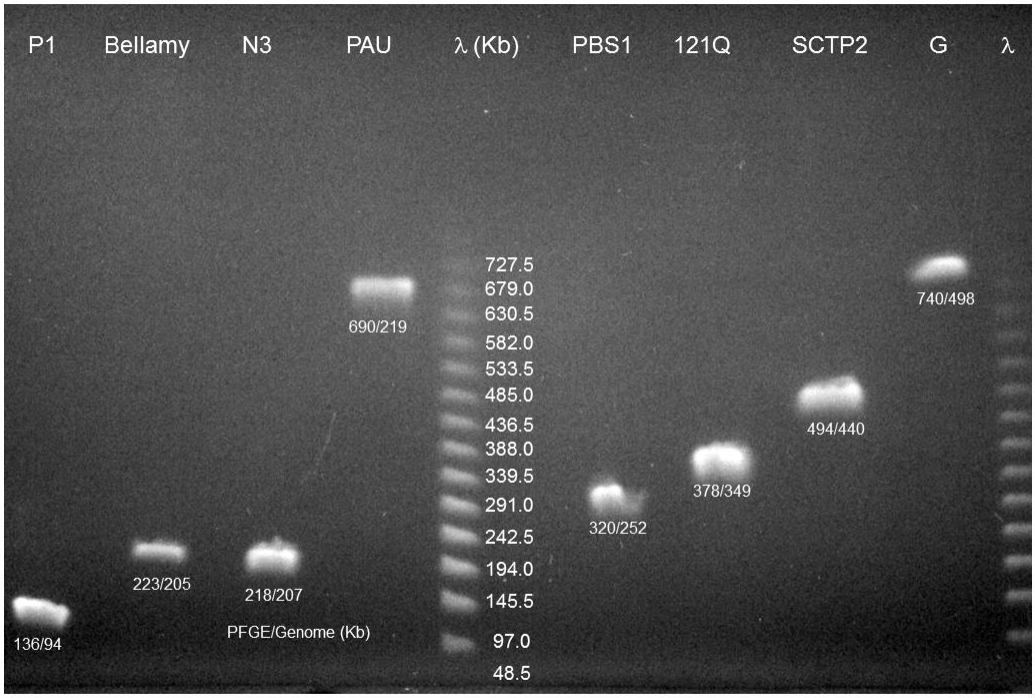
4.3 SIZE OF JUMBO PHAGE CHROMOSOME DNA

The electrophoresis technique for measuring the chromosome size of our jumbo phage is pulsed-field gel electrophoresis (PFGE). By repeatedly switching the direction of the electric field, PFGE overcomes the problem that very large DNA becomes aligned in a constant field and moves in a "reptation" fashion in the agarose matrix and therefore cannot be effectively separated (Schwartz and Cantor 1984). For each sample of PFGE, about 1 X 10^7 phages were mixed with warm 1% agarose and solidified into a gel plug. Incubating the gel plug in a pH 9.5 buffer with EDTA, N-laurylsarcosine and proteinase K in a water bath digested off the protein components of the phage and exposed the DNA. The gel plug was inserted into the well of 1% agarose gel and electrophoresed in a Bio-Rad CHEF-DR II system.

The DNA size of seven jumbo phages was studied by PFGE (Figure 29). Five of them (Bellamy, N3, PBS1, 121Q and SCTP2) show expected DNA packing density and terminal redundancy, according to the apparent size measured from the gel (See 4.4 for details). However, the apparent size of the PAU DNA is 690 Kb, an incredibly larger number compared to its genome size of 219 Kb. Also, the size of the G DNA is 49% larger than its genome size according to the PFGE gel. Since cytosine of PAU and G DNA are modified, the presumed cause of the abnormal migration of PAU and G DNA may be the reduced negative mass/charge ratio by the modifying group.

Figure 29 PFGE of jumbo phage DNA

The DNA of P1 and seven jumbo phages were electrophoresed through a 1% agarose gel in pulsed-field. The position of DNA was organized based on the T-number of the phages. The λ-DNA ladder was loaded in two lanes. The numbers below each DNA band are the apparent size of the band measured from the gel and the genome size of the phage. This exposure of the gel displays bands with high contrasts and low background. However, it does not display the lowest band of the ladder on the bottom edge of the gel, which can be clearly seen in another exposure of the same gel (Figure 30B).



PFGE of jumbo phage DNA

The method for making an accurate estimate of the modified phage DNA is 2D-PFGE, inspired by Dr Phil Serwer at University of Texas Health Science Center at San Antonio. After PFGE in one direction, the phage DNA is electrophoresed in a constant field perpendicular to the original direction. Without pulsing, the migration of the large phage DNA in the second electrophoresis only depends on the charge/mass but not on the size of the DNA. Therefore how much the DNA modification alters the moving speed of a sample DNA in electrophoresis can be quantified by comparing the travel distance of the sample DNA to that of the unmodified reference DNA. This information can be applied on the PFGE to calculate the location of unmodified DNA with the same size as the sample DNA, and therefore determine the size of the sample DNA. Figure 30 illustrates how the size of the PAU DNA is determined by a 2D-PFGE gel a PFGE gel.

The DNA of four jumbo phages (N3, PAU, SCTP2 and G) was prepared in a single gel plug and inserted in the top right well together with a thin piece of λ DNA ladder (Figure 30A). The PFGE performed on the vertical direction separates the DNA based on the mass/charge ratio and the size of the DNA molecules. In the secondary electrophoresis, the separated DNA in the right lane moved to the left in a constant field presumably only based on the mass/charge ratio. The ratio of the travel distance of the sample DNA over the travel distance of the reference DNA is an accurate quantification of how the modification affects the mobility of DNA in the electric field. Because the electric field was not perfectly identical through the whole gel, the travel distance of the DNA varied based on where DNA was at the beginning of the second electrophoresis. This problem is solved by choosing the vertically closest λ ladder band to the sample DNA as the reference DNA. The 2D-PFGE gel provides strong evidence that the cytosine modification decreases the mobility of the PAU and G DNA. In the second

electrophoresis, the travel distances of PAU DNA and G DNA are 74% and 93% of the reference DNA. With no surprise, the mobility of the N3 and SCTP2 DNA is identical to the mobility of the reference DNA.

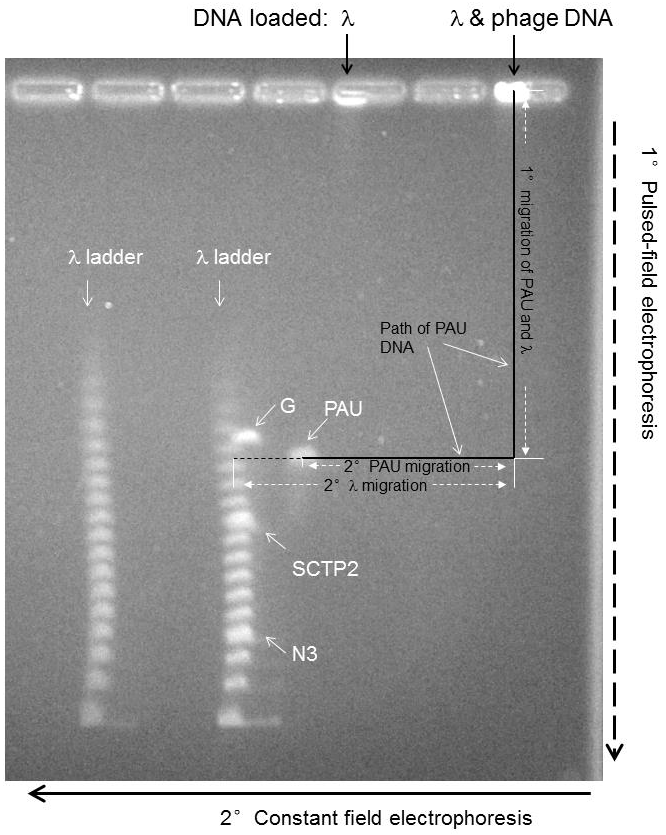
Knowing the relative mobility of PAU and G DNA to their reference DNA, the size of the PAU and G DNA can be calculated on the previous PFGE gel in (Figure 29). Figure 30B demonstrates how the size of the PAU DNA is determined. Assume that an unmodified PAU DNA sample is loaded together with modified PAU DNA in the same well. This unmodified DNA runs faster than the modified DNA, forming a band below the modified PAU DNA band. The distance between the modified DNA band and the loading well will be 74% of the distance between the unmodified DNA band and the well, which determines the location of the unmodified DNA marked by the white box in the figure. The measured size of the assumed unmodified DNA band at the white box, 296 Kb, is the measured size of PAU DNA by the 2D-PFGE method. The size of the G DNA is measured to be 630 Kb using the same approach. These two numbers are much more reliable estimate of the chromosome sizes of PAU and G DNA than the estimates that do not consider DNA modification.

Figure 30 Measurement of the size of PAU DNA by 2D-PFGE

A. A 2D-PFGE gel. The PFGE was performed at the vertical direction, followed by the second electrophoresis at the horizontal direction. The phage DNA and the ladder are labeled. The path of the PAU DNA is indicated by two black lines. The additional distance that the ladder DNA moved is indicated by a dashed black line. All the phage DNA was loaded in the top right well together with the λ ladder. Additional λ ladder was loaded in the third well from the right. The two parallel ladders check if the direction of 2^{nd} electrophoresis was indeed perpendicular to the direction of the 1^{st} electrophoresis.

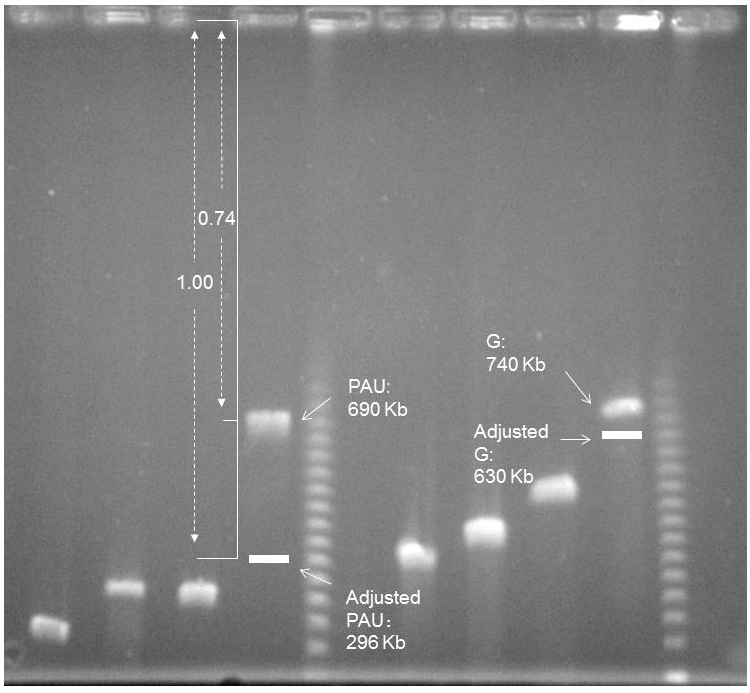
B. A PFGE gel to calculate the size of PAU DNA. The image is a different exposure of the same gel in Figure 29. The position of the imaginary unmodified PAU DNA is determined by the relative mobility of the modified DNA to the reference DNA determined in A. The white boxes indicate the location of the unmodified PAU and G DNA.

A



2D-PFGE of G, PAU, SCTP2 and N3 DNA

B



PAU and G DNA size calculated based on the 2D-PFGE data

4.4 CAPSID DIMENSIONS AND DNA PACKING DENSITY IN PHAGES

The genomic DNA of the tailed phage is efficiently organized in a compact space within the capsid shell to a very high pressure of several tens of atmospheres (Rau, Lee et al. 1984). Jumbo phages may have both advantages and challenges in packing the full sized genome in more spacious capsids compared with the well-studied smaller phages. The DNA size measured by 2D-PFGE combined with the capsid structure determined by cryo-EM reconstruction makes it possible to analyze the relationship of DNA content and capsid space in our jumbo phages. I am curious if the data can provide tangible evidence about the general relationship between DNA packing and capsid size in tailed phages.

4.4.1 Calculation of capsid volume and DNA packing density

The capsid volume based on cryo-EM maps of phage capsids filled with DNA is treated as a convex icosahedron. Let *d* be the vertex to vertex distance of the icosahedral empty space along the five-fold axis, the volume of the icosahedral space (*Vi*) is:

$$Vi \sim 0.32 * d^3$$

For prolate capsids, the total volume (*Vt*) is estimated as the combination of two parts, the volume of an icosahedron (*Vi*) and the volume of a pentagonal prism (*Vp*). Let *a* be the length of the edge that connects two neighboring vertices, *h* be the distance along the five-fold axis that connects the portal to the distal vertex and Δh be the elongated length of the prolate capsid. The total volume of the prolate capsid (*Vt*) is estimated as follows:

$$a \sim 0.53 * d$$

$$\Delta h = h - d$$

$$Vp \sim 1.72 * a^2 * \Delta h$$

$$Vt = Vi + Vp \sim 0.48 * d^2 * h - 0.16 * d^3$$

The DNA packing density of our jumbo phages and P1 were calculated based on my PFGE and cryo-EM data.

DNA packing density, the ratio of DNA size over capsid volume, is an indicator of how efficiently a phage utilizes the capsid space. The DNA packing density of the jumbo phages and P1 were calculated based on my PFGE and cryo-EM data. The presence of internal proteins in these phages raises the question if ignoring the capsid space occupied by internal proteins will significantly underestimate the DNA density and pressure in phages with lots of internal proteins. Take PAU, in which the capsid has plenty of internal proteins, for an example: a rough estimate (Harpaz, Gerstein et al. 1994) of the total volume of the internal proteins as soluble proteins is about 33,000 nm^3, which is about 6% of the total capsid volume. This marginal percentage suggests that it is safe to ignore the volume of the internal protein which is not studied in detail in most phages.

4.4.2 Capsid and DNA size in phages with a range of T-numbers

I did a broad survey on the availability of genome sequence length, terminal redundancy and high resolution capsid structure for calculating DNA packing density in other phages. Several small or mid-sized, well documented phages show up on my radar, including λ, P22, T7, T4 and SPO1. Fortunately, these model phages and the jumbo phages together cover most T-numbers that have been found in tailed phage, except three rare T-numbers of 4, 9 and 12. The T-number, capsid diameter, genome sequence size, chromosome size, terminal redundancy, DNA packing density and the DNA ring spacing observable in the cryo-EM model of these phages are listed in

Table 4. Apparently, the capsid diameters and volume, genome size and chromosome size are positively correlated with the T-number.

The averaging of the densely packed DNA in cryo-EM reconstruction results in ring-shaped DNA spacing commonly observed in cryo-EM model of capsids filled with DNA (Figure 21). DNA ring-spacing is reported to be negatively correlated with DNA packing density in phage or icosahedral virus mutants that package a partial length of genomic DNA (Earnshaw and Harrison 1977; Bhella, Rixon et al. 2000). However, Table 4 shows that DNA ring spacing and DNA packing density in wild-type phages appear not correlated. Correlation tests conducted using Pearson method suggests no significant correlation between DNA packing density and DNA spacing (or DNA spacing to the power of -3), with $p = 0.07$ (or 0.1). Admittedly, the DNA ring spacing data measured from early low resolution cryo-EM models may not be accurate and I only have 10 complete data pairs. More and higher quality data are needed to learn if DNA ring spacing is related to DNA packing density in wild-type phages.

Terminal redundancy varies in different phages from 0% in λ to 45% in P1. The general observation is that terminal redundancy is larger in jumbo phages, with the exception of the T=13 phage P1 and the T=19 phage N3. While a terminal redundancy at 45% seems extreme, the consistency of the DNA packing density of P1 and the other phages suggests that it may not be a complete overestimate. P1's high terminal redundancy and production of smaller phages suggests that it might have experienced a recent evolutionary event that expanded the capsid size. The terminal redundancy of N3 is very small compared with the other jumbo phages. The low terminal redundancy together with a high DNA packing density allow the N3 capsid to accommodate a relatively big genome close to the genome size of the T=25 phage PAU but

much larger than that of the T=16 phage SPO1. Filling up the capsid space with a big genome carrying more beneficial genes may be N3's strategy to survive.

Table 4 Dimensions of phage capsid and DNA

Phage	T/Q	Diameter		Capsid volume $X10^3$ nm^3	Sequence lengths		Redun-dancy	DNA packing density bp/nm^3	DNA ring spacing nm
		Min nm	Max nm		DNA Kb	Genome Kb			
λ	7	52	67	64.4	48.5	48.5	0%	0.75	2.2
P22	7	59	64	58.3	42.5	41.7	2%	0.73	2.1
T7	7	56	66	50.1	39.3	39.3	0%	0.78	2.2
P1	13	81	94	212	136	93.6	45%	0.64	2.4
T4	13/20	86	120	295	175	170	3%	0.59	2.4
Bellamy	13/24	97	135	381	223	205	9%	0.58	-
SPO1	16	87	99	258	146	133	10%	0.55	2.1
N3	19	93	119	376	213	207	3%	0.57	2.3
PAU	25	116	130	560	296	219	35%	0.53	2.5
PBS1	27	121	140	644	320	252	26%	0.50	2.5
121Q	28	120	140	708	378	349	8%	0.53	2.3
SCTP2	39	144	163	1120	495	440	13%	0.44	-
G	52	150	173	1640	626	498	26%	0.38	-

Notes:

1. The maximum diameter is the vertex-to-vertex distance along the five-fold axis; the minimum diameter is the distance along the three-fold axis.
2. The capsid volume is calculated using cryo-EM maps of phage capsids filled with DNA.
3. The DNA ring spacing is the average distance of the nearby DNA layers observed in cryo-EM model. The ring spacing could not be determined in Bellamy, SCTP2 and G because resolution of the models is low or the model is based on empty capsids.
4. Reference: λ (Feiss and Widner 1982; Lander, Evilevitch et al. 2008), P22 (Spanova 1992; Lander, Tang et al. 2006), T7 (Kelly and Thomas 1969; Guo, Liu et al. 2014), T4

(Streisinger, Edgar et al. 1964; Fokine, Chipman et al. 2004) and SPO1 (Cregg and Stewart 1978; Duda, Hendrix et al. 2006)

4.4.3 The relationship between DNA packing density and T-number

T-number is convenient to describe the capsid size. But when studying the relationship between DNA packing density and capsid size, can T-number be used to accurately represent capsid size instead of the less conventional and convenient choice of using the actual capsid volume? It is important to validate this argument because capsid size may vary based on the rich structural diversities in different capsids, not only about the size of primary sequence and accessory domain of MCPs, but also about the contribution of decoration proteins to the continuous capsid shell (i.e., the decoration protein that fills the hole in capsomers of 121Q in 3.3.3). If T-number, which equals the number of MCP subunits in the asymmetric subunit, is the dominant factor in determining capsid volume, it should be proportional to the surface area, and therefore proportional to the capsid volume to the power of 2/3. The plot of volume$^{2/3}$ vs T-number in isometric capsids (Figure 31) shows a strong linear relationship with a near zero intercept (Intercept: -2.1, p-value: $2 * 10^{-12}$, adjusted R-squared: 0.996 in linear regression test). Therefore, the data suggests that T-number is a reliable indicator of capsid volume. The data also imply that the dimensions of the HK97 fold and packing of subunits and capsomers in tailed phages are very similar.

The most interesting observation in Table 4 is that DNA packing density is negatively related to T-number. The DNA packing density in the 13 surveyed phages shows a wide range from 0.38 bp/nm^3 to 0.78 bp/nm^3. The plot of DNA packing density vs T-number displays a steadily trend: larger capsids tend to have lower DNA packing density (Figure 32A). The largest T=52 phage G has the lowest DNA packing density in these phages, less than half of the highest DNA packing density in the T=7 phage T7. How DNA packing density is quantitatively related to T-number is currently unknown. However, there is an approximate linear relationship between

the DNA packing density and the inverse square root of T-number in the 13 surveyed phages (Figure 32B).

For jumbo phages, terminal redundancy appears related to DNA packing density. Compared with their neighbors close in size, both N3 and 121Q have obviously lower terminal redundancy and higher DNA packing density. A possible explanation is that both phages have preferentially acquired exogenous genes. In this process, maintaining a large terminal redundancy, as in other jumbo phages, will lead to a much higher DNA packing density that is not favored. In this case, reducing the terminal redundancy to decrement DNA size becomes a convenient solution.

Figure 31 The linear relationship of volume to the power of 2/3 and T-number in isometric capsids

Only the data of the isometric capsids (blue) are used in fitting the linear regression line (red). Summary of the linear fitting: intercept: -2.1, p-value: $2 * 10^{-12}$, and adjusted R-squared: 0.996.

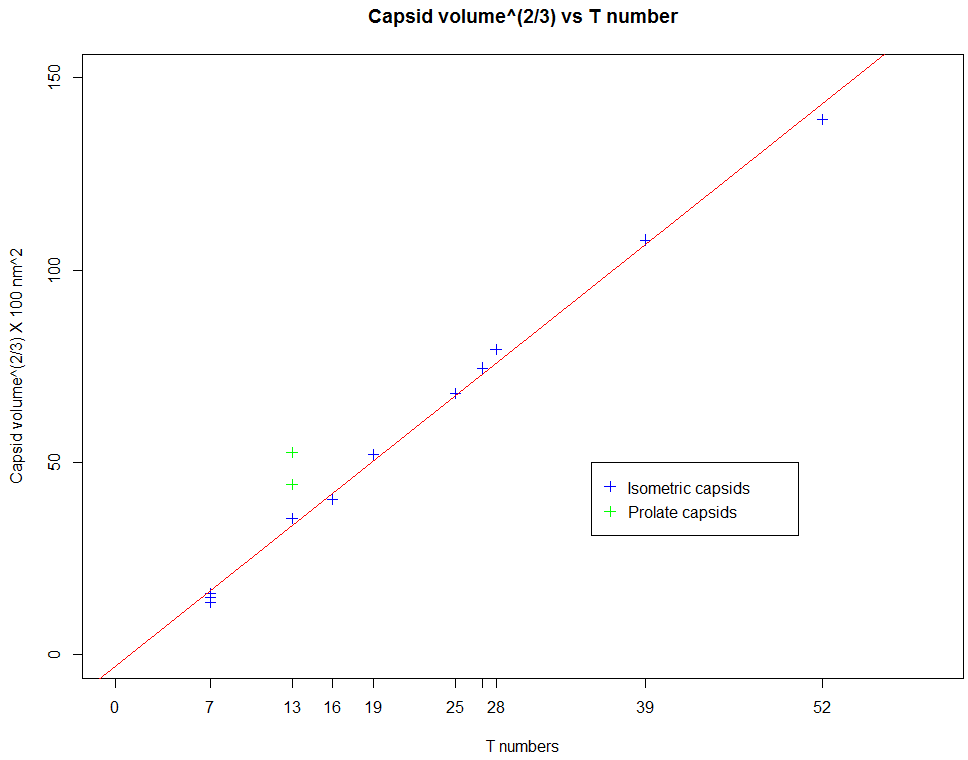
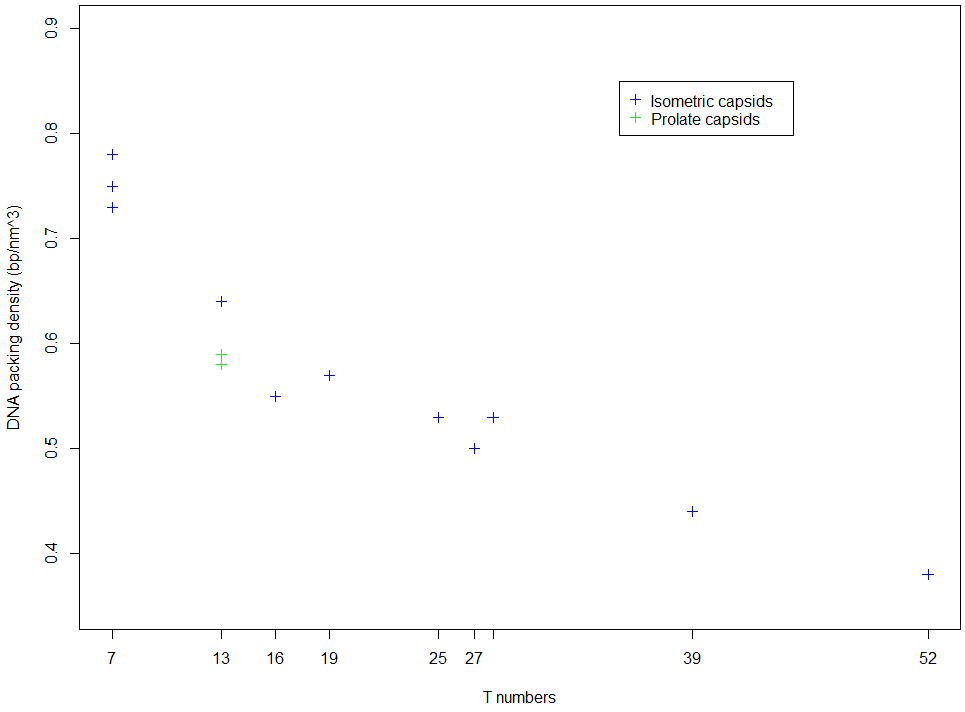


Figure 32 The relationship of DNA packing density to T-number

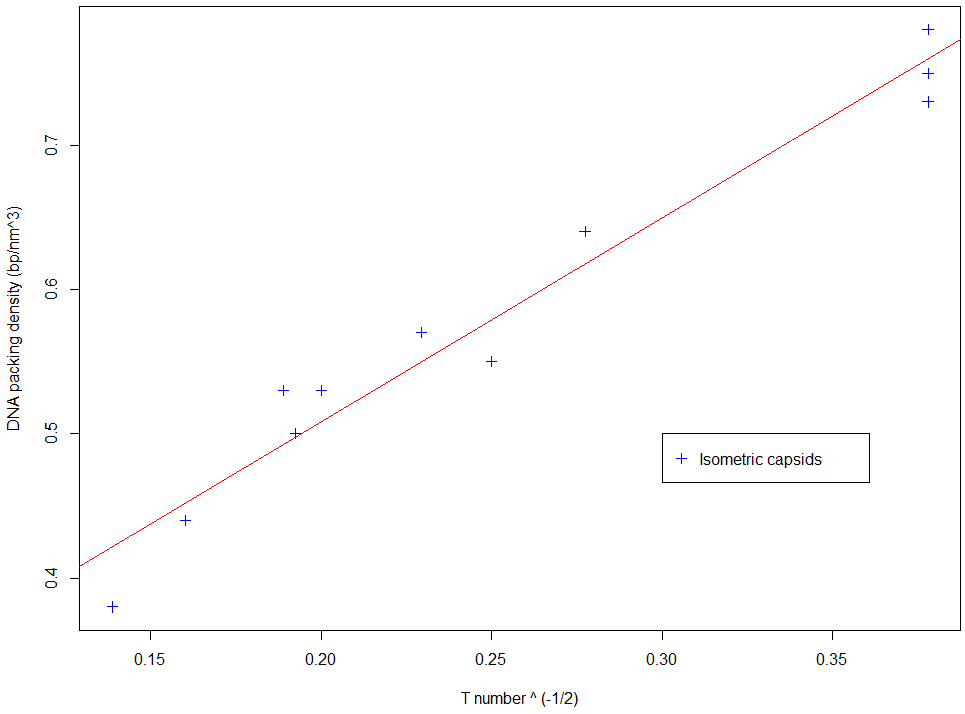
A. A plot of DNA packing density versus T-number.

B. A plot of DNA packing density versus the multiplicative inverse of the square root of T-number. Only data of the isometric phages are displayed. The line of linear regression fitting is drawn in red. Summary of the linear fitting: p-value: $1 * 10^{-7}$, and adjusted R-squared: 0.95.

A


DNA packing density vs T numbers
Isometric capsids
Prolate capsids
DNA packing density (bp/nm^3)
0.4
0.5
0.6
0.7
0.8
0.9
7
13
16
19
25
27
39
52
T numbers

B


DNA packing density vs T number ^ (-1/2)
Isometric capsids
DNA packing density (bp/nm^3)
0.4
0.5
0.6
0.7
0.15
0.20
0.25
0.30
0.35
T number ^ (-1/2)

4.5 DISCUSSION

4.5.1 Cytosine modification in PAU DNA

Since the first DNA base modification, 5-hydroxymethylcytosine (hmCyt), was discovered in T-even bacteriophages in 1953 (Wyatt 1953; Wyatt and Cohen 1953), a variety of modified bases have been found in many bacteriophages (Gommers-Ampt and Borst 1995). A common base can be either completely or partially replaced by a modified base. A major function of base modification in bacteriophages is protecting phage DNA from host nucleases or alternatively from phage-encoded nucleases whose function is to degrade host DNA but not phage DNA. The modified bases can also play some roles in transcription and the replication, packaging and injection of DNA (Warren 1980).

The DNA modification in phage PAU is probably a new kind of cytosine modification yet to be identified. The greatly increased affinity of the modified cytidine to the C18 column suggests that the modifying group may have a large hydrophobic structure. However, the much slower migration of PAU DNA in electrophoresis also suggests that the modifying group may be positively charged. The modified cytosine was unstable during the purification procedure. Oxidation may play a role in the degradation of the modified cytosine sample purified by HPLC. These characteristics in a modified cytosine have not been reported before.

Interestingly, the cytosine in G DNA is also partially modified. Similar to the PAU DNA, the G DNA moves slower during electrophoresis and the modification increases the hydrophobicity of the base. It is conceivable that the modification in PAU and G DNA belong to a same unidentified type of base modification.

4.5.2 DNA size measured by 2D pulsed filed gel electrophoresis (2D-PFGE)

An innovative approach, 2D-PFGE, was used to determine the chromosome size of our jumbo phages. This method is superior to the reported ways for determining the DNA size and terminal redundancy of phage DNA, including:

1. The genetic approach. It requires a clear understanding of the replication and packaging of the phage DNA in a well-studied host, and therefore not practical for a random phage in a random host.
2. Measuring the length of the released DNA imaged by negative stain EM. It is a universal method, but hardly repeatable for big phages because of technical difficulties in sample handling and length measuring.
3. PFGE. It is common method to determine the length of very large DNA, but does not consider DNA modification that may change the behavior of the DNA in electrophoresis.

For measuring the length of large DNA without knowledge of its properties, the 2D-PFGE method takes care of the problems caused by potential base modification and it is as convenient and repeatable as PFGE. Therefore, it is the method of choice for studying the size of DNA from new phages.

4.5.3 Characteristics of DNA in jumbo phages

The survey on the dimension of capsid and DNA in phages with a variety of T-numbers shows that jumbo phages have higher terminal redundancy and lower DNA packing density than

smaller phages. Some percentage of terminal redundancy is necessary for each chromosome molecule to carry intact copies of all genes, but jumbo phages probably have other reasons for keeping a large terminal redundancy of more than 20%. One possible explanation is related to the imprecision of the DNA length in phages with headful DNA packaging. The reported variation in DNA size in P22, Mu and SPPI is about +/- 2% the genome length in laboratory conditions (Chow and Bukhari 1978; Humphreys and Trautner 1981; Casjens and Hayden 1988; Casjens and Gilcrease 2009). Apparently conditions in natural environments are more complicated and are usually not as ideal as in laboratories so the variation in the DNA size is very likely to be larger than +/- 2%. In this case, a large terminal redundancy may reduce the possibility of packaging an incomplete copy of genome into a capsid, at a cost of more capsid space which is less a problem for jumbo phages. Therefore, a large terminal redundancy may benefit jumbo phages more than small phages.

To my knowledge, no previous study has discussed why different phages have different DNA packing density while the major capsid protein, the portal complex protein and the DNA packaging motor are the most conserved proteins in tailed phages. I suggest two possible explanations for the much lower DNA packing density in jumbo phages:

1. Jumbo phages are not able to maintain the DNA packing density as high as in small phages because the capsid has some physical limit. The physical limit is likely imposed by the capsid size, according to the relationship between DNA packing density and T-number. This is my favored hypothesis and I present a simple model in the next section to explain this limit. This hypothesis implies that phages generally package as much DNA as possible to improve fitness.

2. Jumbo phages prefer a low DNA packing density because it increases fitness. The capsid serves as a container for keeping the pressurized DNA in high energy state, which make the phage sensitive to many external factors such as temperature, pH and ions (Jonczyk, Klak et al. 2011). Having a low DNA packing density reduces the internal pressure and thus alleviates the effect of these environmental pressures on the phage (Bauer and Evilevitch 2015; Bauer, Li et al. 2015). Jumbo phages embrace the advantages of low DNA packing density because of the spacious capsid. However, this hypothesis has a problem in that it implies that small phages don't have to have high DNA packing density and jumbo phages don't have to have low DNA packing density, but no such examples were obtained when I did the literature search for preparing Table 4.

4.5.4 A simple model for the relationship between DNA packing density and capsid size

A naïve model is better than no model. I describe here a simple model about why the size of the phage capsid can be the physical limit that determines the DNA packing density. Assumptions are made to simplify the model, including:

1. The capsid is a perfect icosahedron.
2. The pressure caused by the internal DNA is uniform and only depends on the DNA packing density.
3. The capsid shell is infinitely thin.

4. Interactions among MCP subunits that balance the pressure from the DNA are uniform. The interaction is simply considered as a uniform tension in the capsid shell.

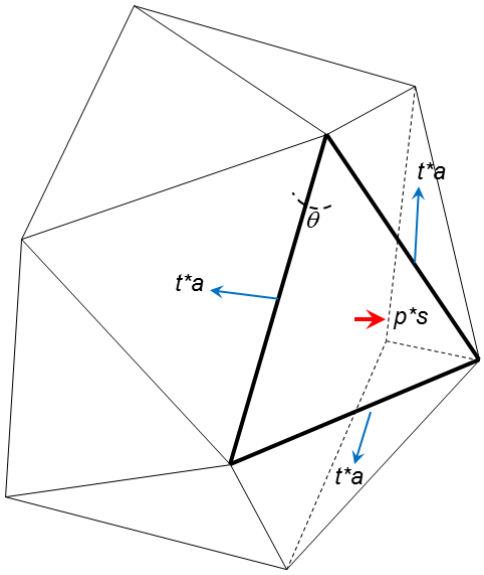
Consider the balanced forces in a facet of the icosahedral capsid (Figure 33). Let p be the pressure of the DNA and S be the area of the triangular facet. The force F_{out} that pushes the facet away from the center of the capsid is:

$$F_{out} = p * S.$$

Let t be the tension in the capsid shell, a be the distance of two nearby vertices and θ be the dihedral angle between two interacting facets. The force F_{in} that pulls the facet toward to the center of the capsid is:

$$F_{in} = 3 * t * a * cos(\pi - \theta).$$

Figure 33 The force in a triangular facet



In a stable capsid,

$$F_{out} = F_{in}.$$

Apparently, S is proportional to the T-number (T) and the square of a:

$$S \propto T,$$

$$S \propto a * a.$$

Combine the above equations together and get:

$$t = C_0 * p * \sqrt{T}, \qquad (1)$$

where C_0 is a constant.

Equation 1 suggests that the interaction between the MCP subunit that overcomes the pressure from the DNA is determined not only by the pressure, but also by the size of the capsid. Stronger interactions are needed in a jumbo phage if it maintains the same DNA packing density observed in small phages. For the T=52 phage G, the inter-molecular interaction of its MCP has to be more than 2 times stronger than that in the T=7 phage λ.

Now I make the last assumption that all phages package as much DNA as possible to maximize the amount of genetic information it can hold. In this case the tension in the capsid shell t is at a point that is limited by the strength of the inter-molecular interaction of the MCP. Because the HK97 fold is highly conserved, the tension in the capsid shell t is approximately a constant in phages of all sizes. From equation 1, I can get:

$$p = C_t / \sqrt{T}, \qquad (2)$$

where C_t is a constant related to t. Equation 2 shows that the pressure of the DNA is linear to the inverse of the square root of the T-number if the DNA packing density is limited by the strength of the inter-molecular interaction of the MCP. It is consistent with the negative correlation between the DNA packing density and the capsid size discussed in 4.4.3.

The quantitative relationship between the pressure and the packing density of the DNA in phages is not well understood. Studies on the force during DNA packaging suggests that the force required to overcome the internal pressure as a function of percent genome packed is approximately exponential, and is roughly linear for between 50~80% of the genome packed in ϕ29 and T7 (Smith, Tans et al. 2001; Purohit, Kondev et al. 2003; Locker, Fuller et al. 2007). The 50~80% genome packed is equivalent to an interval of DNA packing density of 0.39~0.62 bp/nm^3. Because the DNA packing densities of all T>7 phages are very close or within this interval, the DNA pressure *p* of most phages in Table 4 is approximately proportional to the DNA packing density. Let *D* be the DNA packing density. For phages with T>7, *p* can be replaced by *D*, so

$$D = C / \sqrt{T}, \qquad (3)$$

where *C* is a constant. Equation 3 is consistent to the result that DNA packing density is approximately proportional to the inverse square root of T-number (Figure 32B).

In summary, the above model based on the current understanding of the capsid and DNA in tailed phages is consistent with the data on the relationship between DNA packing density and T-number. The model may explain why most jumbo phages, if not all, have a number of decoration proteins: the decoration protein may strengthen the capsid shell and push the limit of the DNA packing density up a little bit. The model also suggests that expanding the size as a strategy of increasing the amount of genetic information that a capsid can carry may receive diminishing benefits as the size of the capsid grows, a possible reason why small phages prevail in natural environments and why we haven't found capsids bigger than T=52.

5.0 DISCUSSION

5.1 DIFFERENCES AND SIMILARITIES BETWEEN THE PAU AND THE 121Q CAPSIDS

5.1.1 The capsid shell

As we expected, the capsid shell of both phages is formed by MCPs with the HK97 fold. The PAU MCP is larger than the T4 vertex protein gp24 and its N-terminus forms the extra knob density on the outer surface. The size of the 121Q MCP is very close to gp24. However, the 121Q MCP does not form the entire capsid shell but leaves holes in the center of the hexamer and near the local 2-fold symmetry axis between two neighboring capsomers. Large holes between the HK97 fold have not been reported in other phage capsids, and therefore this feature of the 121Q capsid is worth further investigation.

5.1.2 Decoration on the outer surface

We show that all the decoration densities on the outer surface of the PAU and 121Q capsids are different. The PAU capsid only contains the banana density on the hexamers surrounding the pentamer. The interaction between the banana protein and the capsid is not strong. However, the banana density introduces conformational changes to the MCP subunit that it interacts with,

stabilizing the interaction between the hexamers with the banana structure. The 121Q capsid contains three types of extra densities on the outer surface, the hair-like head fiber at the global 2-fold symmetry axis, the dimeric density around the local 2-fold symmetry axis and the spherical density at the center of the hexamer. The head fiber seems not essential to the capsid structure but showed high affinity for the host debris. The dimeric density and the spherical density may be essential to the capsid structure because they fill up the holes on the capsid shell left by the MCP.

5.1.3 Decoration on the inner surface

Both capsids contain the mango density, a pentameric density attached to the inner surface of the pentamer. Two observations about the mango density suggest that it may be important to the structural integrity of the capsid shell. First, our urea extraction experiment shows that the interaction between the mango density and the pentamer is strong. Second, the mango density may be a conserved feature in isometric jumbo capsids. Our cryo-EM models of the T=52 phage G capsid and the T=19 phage N3 capsid also contain the mango density. In addition to our jumbo phage models, we observe that, although the authors of the papers did not specify it, the two T=27 capsids of phiKZ (Fokine, Kostyuchenko et al. 2005) and phiRSL1(Effantin, Hamasaki et al. 2013) also contain extra protein structure under the pentamer. The presence of the mango density in all six jumbo phages implies that it may play some important roles in capsid assembly, which can be explained as a consequence of selecting the MCP for forming the flat hexamer but not the curved pentamer. As a result of this selection, the MCP forms more stable hexamers, which outnumber the pentamer in jumbo phages, and the less stable pentamer. With a relatively

small cost of the mango protein that stabilizes the twelve pentamers, the strength of the entire capsid may be significantly improved.

5.1.4 Structure formed by internal proteins

The PAU capsid contains at least seven internal proteins and the 121Q capsid contains more. The PAU internal proteins form the irregular patch in the empty capsid, but no bubblegram was observed after heavy beam exposure during cryo-EM imaging. The 121Q phage shows bubblegram after heavy beam exposure. We do not know if the internal protein of the 121Q empty capsid forms a structure that is visible during cryo-EM imaging.

5.2 FUTURE WORK ON JUMBO CAPSIDS

5.2.1 Future work on PAU and 121Q capsids

I think at least two aspects in the PAU project are worth further investigation. One is to improve the cryo-EM model of the urea treated capsid to better than 10Å, which may show how the presence of the banana affects the conformation of the hexamer it interacts with. The other aspect is about the high resolution cryo-EM structure and biochemistry of the heated capsid which loses the banana protein and the major mango protein. The structure and biochemistry of the capsid without the mango protein may shed some light on the role of the mango structure. The above two objects will probably be reached without any technical problem, considering the easiness of PAU capsid purification and the advance of the cryo-EM techniques in the recent years.

However, it is difficult to work with the genetics of PAU because the genetic background of *Sphingomonas paucimobilis* is poorly understood.

For the 121Q capsid, the holes left by the HK97 fold and the extra surface densities that fill the holes are very unusual and interesting. Experiments similar to the urea extraction and cryo-EM of PAU empty capsids can be carried out to study the protein identity, structure and function of the extra densities, providing that a pure sample of empty capsids can be isolated. The hair-like head fiber and sheath fiber that have high affinity for the host membrane bring difficulties to the purification of 121Q particles. A possible solution is studying the 121Q-like phages, instead of 121Q. The MCP and the two potential proteins that form the dimeric density, gp599 and gp600, are highly conserved in this group of jumbo phages, but the head fiber protein gp552 is much less conserved. The vB_KleM-RaK2 capsid may not have the head fiber because BLAST of gp552 shows that RaK2 does not contain a protein similar to gp552, and therefore the Rak2 capsid may be easier to purify.

The head and sheath fiber protein gp552 (and potentially the 360 kDa protein gp563) is an interesting protein. In the future one may solve the crystal structure of gp552 and answer the question if gp552 is similar to other phage tail fibers in high order structure or is a new kind of fiber protein.

5.2.2 The next jumbo phage to study

There are at least three strategies to extend the jumbo phage project. The first is what we have been working on, finding jumbo phages with potential novel T-numbers, sequencing the genomes and studying the structural features of the capsids. This may be the fastest way to expand our knowledge base about jumbo phages.

The second strategy is to focus on a group of closely related jumbo phages, which may provide more details about the genetics and structure of jumbo phages. For example, the similar genome sequence between the group members may allow sequence analysis on many less conserved genes instead of few conserved structural genes, making it possible to learn how jumbo phages manage alien genes acquired by horizontal gene transfer and paralogs created by gene duplication. Comparing the capsid structure of different members may also help us understand the role of the decoration density, since a structural feature important to capsid assembly is likely to be conserved. I think the four 121Q-like phages are suitable for such a project. And more 121Q-like phages may be isolated from sewage worldwide in the future.

The other strategy is to work on the genetic detail of a jumbo phage. After all, this may be the only way to thoroughly understand the function of the nonconserved genes that fill more than half of the jumbo phage genome. However, there are not many jumbo phages suitable for genetics. Most jumbo phages, as far as we know, are lytic phages, and the hosts of many jumbo phages have limited genetic tools we can use. To my knowledge, the best subject for such a study is the T=27 *Bacillus subtilis* phage PBS1, which is temperate (Takahashi 1961) and easy to grow and purify.

5.3 DNA PACKING IN JUMBO PHAGES

We developed 2D-PFGE, a convenient and repeatable method to determine the length of large DNA regardless of DNA modification. The packing density and terminal redundancy of the jumbo phage DNAs were calculated based on the length of the DNA, the length of the genome sequence and the volume of the capsid structure. In general, the terminal redundancy in jumbo

phages is larger than the terminal redundancy in small and mid-sized phages. The data show a negative relationship between the packing density and the capsid size. We explain this relationship by a model which proposes that the strength limit between the capsomers may be the reason why a jumbo capsid cannot pack DNA as densely as a smaller capsid. More data points can be obtained to verify the negative relationship between the packing density and the capsid size, but right now we do not know how to experimentally test the model.

Our data suggest that the DNA ring spacing observed in the high resolution cryo-EM structure is not related to the DNA packing density, which is contrary to what we would intuitively expect and what was previously reported (Earnshaw and Harrison 1977; Bhella, Rixon et al. 2000). Precise measurements of DNA ring spacing from more phage structures are needed to increase the predicting power of the correlation test between the packing density and the ring spacing.

5.4 THE MODIFIED CYTOSINE IN PAU DNA

We show that migration of PAU DNA in electrophoresis is slowed by the partial modification of the cytosine. The cytosine modification also greatly increases the affinity of the modified cytidine to the C18 column. We suspect that the modifying group is hydrophobic in pH 5.1, the pH of buffer A in the HPLC experiment, and also a positively charged group with a pKa between pH 5.1 and pH 8.3, the pH of 0.5X TBE buffer in PFGE. Unfortunately, the modified cytosine seemed unstable and the identification using mass spectrometry was unsuccessful.

For identification of the modified cytosine in the future, it is better to coordinate with the mass spectrometry facility in advance and work really fast. Dr Andrew Van Demark suggested using NMR if a few micrograms of pure modified cytidine can be obtained.

6.0 MATERIALS AND METHODS

6.1 STRAINS

All the phage and host strains are from the strain collections of the Hendrix lab and the Genome Center.

1) N3: *Sinorhizobium meliloti* RM1021;
2) PAU: *Shingomonas paucimobilis* Uw101;
3) PBS1: *Bacillus subtillis* SB19E;
4) 121Q: *Escherichia coli* MUL-B70.1.

6.2 MEDIA AND BUFFERS

6.2.1 Media

1) Luria Broth (LB)

1% (w/v) tryptone (Difco), 0.5% (w/v) yeast extract (Difco) and 0.5% (w/v) NaCl in DI water, autoclaved for 25min.

2) LB agar

LB with 1.5% (w/v) agar, autoclaved for 25min.

3) Soft agarose for jumbo phage plates

LB and 0.15~0.25% (w/v) agarose in DI water, autoclaved for 25min. $MgSO_4$ was added to 5 mM and $CaCl_2$ was added to 1 mM before the soft agarose was poured to the LB plate.

6.2.2 Buffers

1) Jumbo phage dilution buffer

   10 mM Tris-HCl pH 7.5, 5 mM $MgSO_4$ and 1 mM $CaCl_2$ in DI water, 0.2 μm filtered.

2) Electrophoresis buffer

   50 X TAE: 1 M Tris-acetate, 50 mM EDTA pH 8.0;

   4 X lower SDS buffer: 1.5 M Tris-HCl pH 8.8, 0.4% SDS;

   4 X upper SDS buffer: 0.5 M Tris-HCl pH 6.8, 0.4% SDS;

   10 X SDS Running buffer: 0.25 M Tris base, 2.5 M glycine, 1% SDS;

   4 X SDS sample buffer : 0.25 M Tris-HCl pH 6.8, 40% glycerol, 20% β-mercaptoethanol, 8% SDS;

   5 X TBE buffer: 0.5 M Tris-HCl, 0.5 M boric Acid, 10 mM EDTA, pH 8.3.

3) Nucleoside HPLC

   Buffer A: 2.5%v/v methanol, 0.01 M $NH_4H_2PO_4$, pH 5.3;

   Buffer B: 8% v/v methanol, 0.01 M $NH_4H_2PO_4$, pH 5.1;

   Flushing buffer: 70% v/v methanol.

6.3 METHODS

6.3.1 Phage stock preparation

High titer phage stocks of N3, PAU, PBS1 and 121Q were produced by mixing approximately 10,000 plaque forming units of the phage, 200 uL overnight host cells growing in LB and 10 mL warm soft agarose. The mixture was poured onto a 150 mm × 150 mm fresh LB agar plate and incubated overnight at 30 ℃ for N3 and PAU, and 37 ℃ for PBS1 and 121Q. Phages were extracted from the top agarose by mixing it with an equal volume of phage dilution buffer, followed by vigorous vortexing. The insoluble material was removed by centrifugation in a Beckman JA-50.250 rotor at 8,000 rpm for 10 min. Phages in the supernatant were pelleted in a Beckman Type 45Ti rotor at 4 ℃, 35,000 rpm for 1 hour, and then suspended in buffer.

6.3.2 Particle purification

Phages PAU and PBS1 were further purified using CsCl step gradients (5 mL sample: 3 mL 30% glycerol: 2 mL 1.4 g/mL CsCl: 2 mL 1.6 g/mL CsCl, all layers in dilution buffer) in a Beckman SW41 rotor at 30,000 rpm, 18 ℃ for 90 minutes. Phages N3 and 121Q are sensitive to concentrated CsCl solution, so they were purified by velocity sedimentation in 10% - 45% sucrose gradients in dilution buffer in a Beckman SW41 rotor at 30,000 rpm, 18 ℃ for 60 minutes. The visible phage bands in both gradient types were collected and dialyzed against dilution buffer overnight. The sucrose gradients were repeated for 121Q particles.

6.3.3 Capsid manipulation

Purified jumbo phage particles were mixed with chemical reagents (Urea, guanidine chloride, etc.) and incubated on bench for 1 hour, or heated on desired temperatures in a thermal block for 10 minutes. The sample was diluted to 1 ml and centrifuged in a Beckman Ti80 rotor at 40krpm, 4˚C for 1.5 hour. The supernatant was carefully and thoroughly removed and the pellet was resuspended in dilution buffer.

6.3.4 Cryo-electron microscopy and image reconstruction

Cryo-EM data were acquired by an FEI Tecnai F20 microscope operating at 200kV, or an FEI Polara microscope operating at 300kV. Micrographs were recorded at 50,000 X magnification on a Gatan UltraScan 4000 CCD with an effective pixel size of 2.14 Å. Particles were selected with x3dpreprocess (Conway and Steven 1999) and defocus was estimated with BSOFT (Heymann 2001). Capsid reconstruction was done using AUTO3DEM (Yan, Sinkovits et al. 2007). Tail reconstruction was done by spider (Shaikh, Gao et al. 2008).

6.3.5 Cryo-EM density map manipulation

Visualization of cryo-EM density map and fitting of HK97 gp5 structure in the map were done in UCSF Chimera (Pettersen, Goddard et al. 2004). Visualization of the decoration density was done by zoning and selection of the density around the fitted gp5 coordinates or manually placed beads.

6.3.6 Protein identification by mass spectrometry

121Q protein samples were briefly electrophoresed into the lower SDS gel and the electrophoresis was stopped. The whole band near the interface between the upper gel and lower gel was cut out and sent for a single LC-MS/MS experiment.

6.3.7 Pulsed-field gel electrophoresis (PFGE)

PFGE experiments were conducted on a Bio-Rad CHEF-DR III Pulsed Field Electrophoresis System, according to manufacturer's instructions. Phage chromosome DNA samples were embedded in agarose by combining 1 X 10^7 phages with warm agarose (1%) in water to a total volume of 60 µL and then transferring the mixture into plug mods to solidify. The phage-agarose plugs were soaked in 0.5 mL buffer containing 0.5 M EDTA, 0.01 M Tris-HCl pH 9.5, 1% N-laurylsarcosine and 400 ug/mL proteinase K and incubated in a 60 ℃ water bath overnight. The phage-agarose plugs were soaked in 0.5 X TBE buffer for 1 hour before being inserted into the wells of a 1% agarose gel. Pulsed-field electrophoresis was carried out in 0.5 X TBE buffer, 14 ℃, 6 V/cm voltage, 50 – 90 second switch time, 22 hour running time, 120 ° angle, and using NEB lambda ladder marker.

In the 2D-PFGE experiment, the DNA of G, PAU, SCTP2 and N3, and the lambda marker were mixed in a single gel plug. The gel plug was loaded in the right well and the DNA mixture was electrophoresed in a pulsed-field using the previous PFGE settings. A second electrophoresis without pulsing was then applied perpendicular to the initial PFGE in a Hoefer HE99X large agarose gel box in 0.5 X TBE at 110 V for 8 hours.

6.3.8 Nucleoside identification by HPLC

The nucleoside identification and purification was conducted based on the method introduced in (Kuo, McCune et al. 1980). About 10 uL of 1000 ng/uL phage DNA prepared by phenol extraction was first digested with 10 u/mL DNase I in jumbo phage dilution buffer at 37 ℃ overnight. Two volumes of 30 mM NaAc pH5.2, 2% volume of 50 mM ZnSO4 and 5 units of Nuclease P1 preheated at 65 ℃ for 10 minutes were added to the reaction and mixture was incubated at 37 ℃ overnight. The reaction was mixed with 3 units of bacterium alkaline phosphatase preheated at 95 ℃ for 10 minutes and then was incubated at 37 ℃ overnight.

The solvent was delivered by a Waters automatic sample injection model and monitored by a Waters UV detector between 220 ~ 340 nm. Separation of nucleosides was performed on a reversed-phage C18 column, typically a 300 X 8 mm column. A typical run contained 30 minutes of buffer A and 40 minutes of buffer B at a flow rate of 1.0 mL/min, after the DNA was injected. The temperature of the C18 column was kept at 45 ℃. The flushing buffer was used for 20 minutes after the run and before a 20 minutes re-equilibration with buffer A.

APPENDIX A

PURIFICATION AND STRUCTURE OF BACTERIOPHAGE P1 PROHEAD

A.1 BACKGROUND

Bacteriophage P1 is a model temperate phage infecting a wide range of enteric bacteria (Lobocka, Rose et al. 2004). In the lysogenic cycle, the P1 prophage is maintained as a circular plasmid with one copy per bacterium. P1 has been used as a generalized transducing phage for moving large pieces of DNA between in *Escherichia coli* (Lennox 1955). P1 naturally produces phages with identical tails but capsids with three different sizes, but only the phage with the largest capsid is viable (Walker and Anderson 1970). Two dispensable internal proteins, DarA and DarB, which are injected into the host during infection for protecting phage DNA against host restriction, reduce assembly of the non-viable capsids (Iida, Streiff et al. 1987).

We believe that P1 may be an interesting system to understand how the capsid size is determined in tailed phages. In P1, DarA and DarB may serve as semi-scaffolding proteins that coordinate the major capsid protein in the assembly of the big viable capsid. In my MS thesis project, the P1 phages with three different capsids were isolated; the structure of the three capsids, the T=4 P1m, T=7 P1s and T=13 P1b capsids, were studied by cryo-EM; and the capsid proteins in the mature P1b capsid were identified by N-terminal sequencing. As a follow-up, I

worked on the purification and the structure of P1 procapsids since the procapsid provides the most direct information on how the decision of capsid size is made during capsid assembly. Unfortunately, the P1 procapsids were difficult to purify and unstable in the conditions used, but the progress I made on this project is described here in the hope that if may be useful to a future member of the lab.

A.2 ISOLATION OF A TEMPERATURE SENSITIVE P LYSOGEN WITH PROTEASE DEFECT

Three methods were used to produce P1 procapsids. First, the P1 major capsid protein gp23 was expressed in *Escherichia coli*, using a similar protocol to the expression of H97 gp5. It was quickly abandoned because the expression level of gp23 was very low. Second, we tried to recover several P1 strains with amber mutations in the protease gene from Jean Walker' P1 strain collection stored in our lab. However, the strains seem no longer active after being stored in the freezer for decades and none of them were successfully recovered. Third, we used the *in vivo* λ Red recombineering method described in (Oppenheim, Rattray et al. 2004) to isolate a temperature sensitive P1 lysogen with the protease gene knocked out. Briefly, HME6 cells in early-exponential phase were mixed with P1*cmclr*100 phages at a multiplicity of infection of 3 phages/cell and incubated on bench for 10 minutes for adsorption. The culture was warmed to 42 ˚C for 15 minutes to activate the λ Red functions. The cells were then chilled and made competent for electroporation by repeated washing with cold water and centrifugation in cold. PCR products containing the kanamycin resistance gene flanked by sequences homologous to sequence to the P1 protease region were transformed into the cells. The cells were diluted in LB

broth and incubated on a shaker at 39 ˚C for 30 min to allow the lytic infection of P1*cmclr*100 to finish. The cell free lysate was mixed with LE392 cells and plated on a LB plate with Kan and Cm selection. The plate was incubated at 30 ˚C overnight. Colonies that grew on the plate were P1cmclr100 lysogens with the protease gene replaced by the Kan resistant gene. The mutation was verified by PCR.

A.3 PURIFICATION OF P1 PROCAPSID

P1 procapsids were produced by inducing the temperature sensitive P1cmclr100 lysogen with the protease gene knocked out. The lysogen culture (1 L in a 2.5 L flask) in mid-exponential phase (O.D.550 ~ 0.3) was quickly heated to 42 ℃ and then incubated on a shaker at 37 ℃. The cells were pelleted and lysed in 10 ml procapsid purification buffer (0.15 M phosphate at pH 6.0, 0.06 M NaCl, 5 mM $MgSO_4$, 0.5 mM $CaCl_2$, 0.5 ug/mL DNase I and 10% glycerol). The lysate was centrifuged to remove cell debris and the supernatant that contained the procapsids was collected.

Centrifuging the supernatant through 15% - 45% sucrose gradients in procapsid purification buffer in a SW41 rotor for 36 krpm, 90 min at 18 ℃ generated two P1 bands, a tail band at about 1/3 depth and a procapsid band at about 2/3 depth. The SDS gel of the procapsid band shows a heavy band with the expected size of the unprocessed gp23 (Figure 34A). The negatively stained EM image of the procapsid bands shows that the sample contained both P1b (~80%) and P1s (~20%) capsids (Figure 34B).

The difficulties encountered in purifying the P1 procapsids were many. The P1 procapsid showed high binding affinity to host membranes. Negative stain EM images of the procapsid

often showed particles aggregated with some host membranes. The procapsid also precipitated when the supernatant or the sucrose band was stored in cold over 24 hours. The procapsid also appeared to be very unstable. Many particles under cryo-EM appeared broken (Figure 35A). Many conditions were tested in order to improve the purification or reduce particle damage, including different kinds of buffer, different pHs, using detergents in the buffer and replacing sucrose with odixanol or glycerol. Unfortunately, none of them greatly improve the purity of the sample or stabilize the procapsid (Table 5).

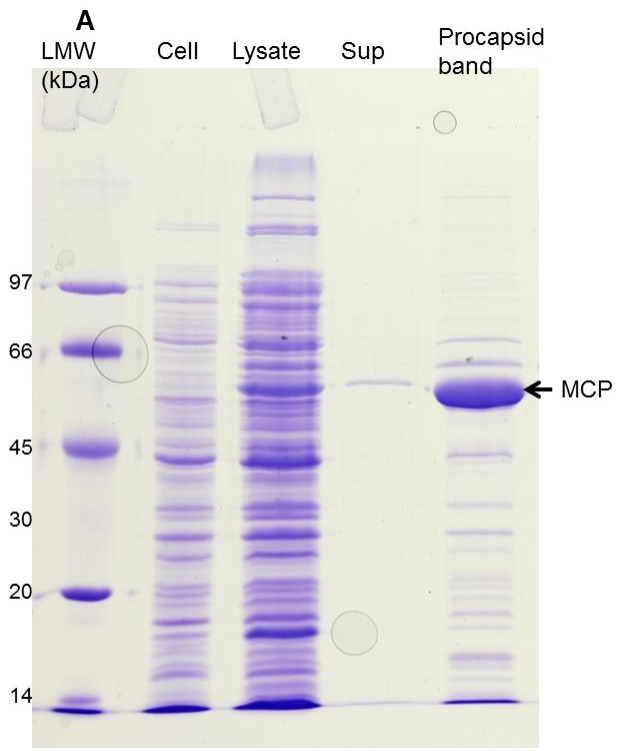
Table 5 List of conditions used in the purification of P1 procapsids

Buffer	**Reasons to use it**	**Results**
Procapsid purification buffer (see above)	Used in purifying T4 procapsids (Traub, Keller et al. 1984)	Used in the project.
10 mM phosphate buffer pH 6.0, 10 mM $MgSO_4$	Low salt	May reduce the solubility of procapsids
50 mM TrisHCl, pH 8.0, 50 mM NaCl, 1 mM EDTA	Used in purifying phi29 procapsids (Tao, Olson et al. 1998); and HK97, T5 and P22 procapsids were purified in TrisHCl buffer, pH 7.5 ~ 8.0	No particles.
Detergents		
1% Triton X-100 in buffer	To solubilize membranes	No improvement.
1% CHAPS in buffer	To solubilize membranes	Improves solubility but the procapsids still aggregated with cell debris under negative stain EM. May cause thick ice and induce procapsid-carbon membrane interaction
1% octylglucoside in buffer	To solubilize membranes	No better than CHAPS
Gradients		
Sucrose gradients		Used in this project. May induce procapsid damage.
Glycerol gradients		Not powerful enough to separate procapsids from membranes.
CsCl step gradients		No procapsids
Iodixanol gradients		Proves better separation than sucrose. But we did not get significantly better cryo-EM data with iodixanol.
Other		
Inducing with higher cell density		Reduces procapsid production
DEAE column	To filter cell debris	No improvement.
Multiple rounds of sucrose gradients		Each additional round reduce some cell debris, at the cost of losing procapsids.
All in cold		No improvement

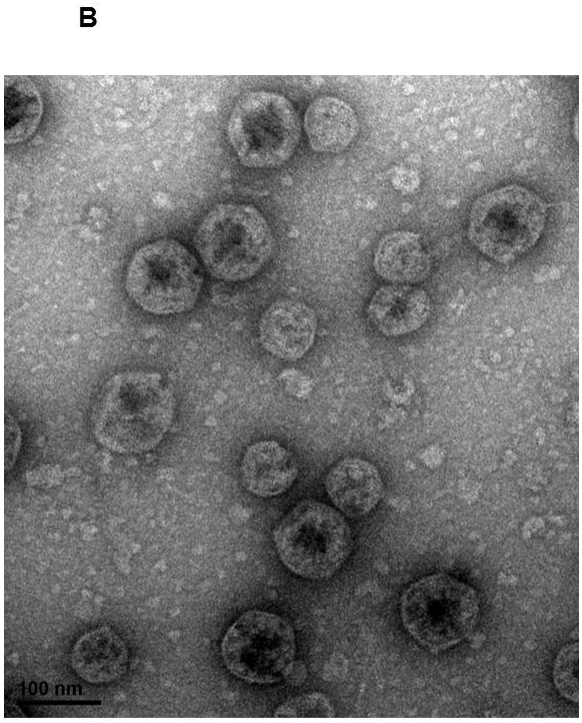
Figure 34 The SDS gel and negative stain EM image of the P1 procapsid sample

A. SDS gel of the P1 procapsid samples. Cell: LE392 cell protein; lysate: whole lysate protein after heat induction of the defective phage; sup: the supernatant of the lysate; and procapsid band: the sucrose band of the procapsid.

B. A negative stain EM picture of the P1 procasid band. The small particle is the P1s capsid and the big particle is the P1b capsid. This sample was treated with CHAPS to reduce the aggregation caused by the cell membrane.



SDS gel of the P1 procapsid sample



A negative EM image of the procapsid band

A.4 CRYO-EM STRUCTURE OF P1 PROCAPSID

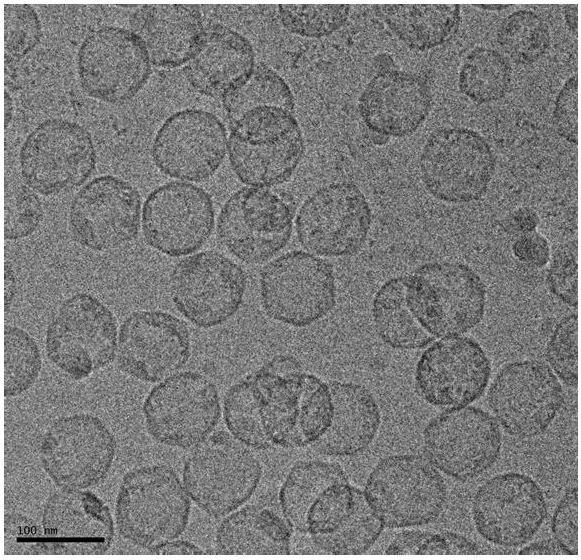
We obtained cryo-EM structure of the P1 procapsid through collaboration with Dr. James Conway and Dr. Alexis Huet. The resolution of the structure is low, mainly because many procapsids in the cryo-EM image appear damaged (Figure 35A). The percentage of the smaller capsids in cryo-EM images is much lower than in the negatively stained EM image. The P1b procapsid appears very similar to the mature P1b capsid in both size and shape, suggesting that it has expanded or matured. The size of the P1s procapsid is in between the sizes of mature P1s and P1m capsids and the hexamers of the P1s procapsid are difficult to recognize. Therefore, the P1s procapsid may be fully expanded but the P1m procapsid might be stuck in some intermediate state. It is not clear why the P1 procapsid is expanded.

Figure 35 Cryo-EM image and structure of P1 procapsids

A. A cryo-EM image of P1 procapsids. Many capsids appear damaged. For unknown reason, the percentage of P1s procapsid is much smaller in the cryo-EM image compared to the negatively stained EM image.

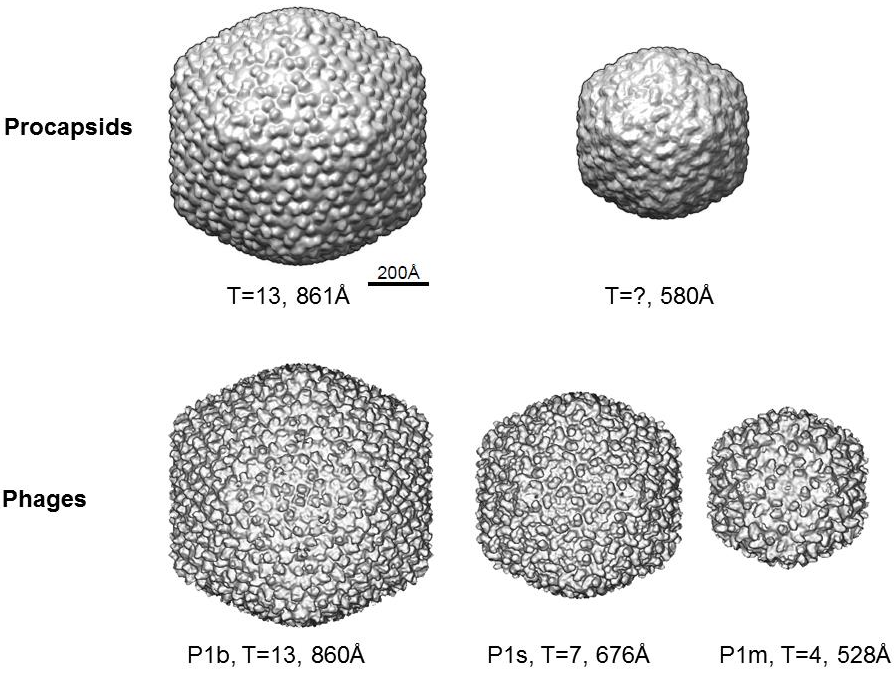
B. Comparison of the cryo-EM structure of procapsids and the mature capsids. The data for the mature capsids were from my M.S. thesis project.

A



A cryo-EM image of the procapsid band

B



Cryo-EM structure of the two P1 procapsids and three P1 mature capsids

A.5 PROTEIN IDENTIFICATION OF P1 PROCAPSID

The protein in the procapsid sample was identified by mass spectrometry (Table 6). Three capsid proteins were detected, including the MCP gp23, DarA and DarB. However, the portal protein was not detected. The experiment neither identified gp21 and ddrB which are internal proteins with about 120 and 50 copies/capsid in the mature P1b capsid (Hua 2010). The data suggest that the protease may be important in recruiting the portal, gp21 and ddrB proteins during P1 procapsid assembly.

Table 6 P1 proteins in the procapsid sample identified by mass spectrometry

Proteins	MW (kDa)	Coverage	Function
Gp23	62	81%	MCP
bplB	19	59%	Putative baseplate protein
Gp22	57	55%	Sheath
LydB	17	22%	LydA(holin) antagonist
UpfA	15	20%	Unkown
bplA	54	19%	Putative baseplate protein
S protein	105	15%	Tail fiber
Mat	29	9.2%	Particle maturation
R protein	16	9.0%	Tail fiber
gp24	29	8.8%	Baseplate protein
DarA	69	5.6%	DarA, defense against restriction A
DarB	252	2.2%	DarB, defense against restriction B

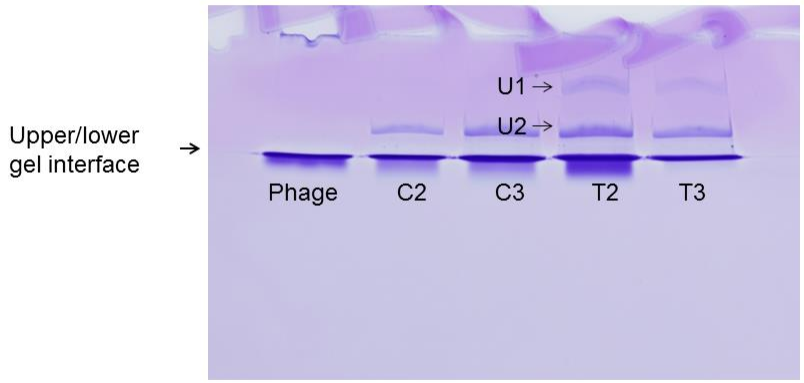
Note: capsid proteins are in bold text.

APPENDIX B

MASS SPECTROMETRY OF 121Q PROTEIN

The protein was briefly electrophoresed into the lower SDS gel. All the proteins were located near the upper/lower gel interface except two proteins gp563 (U1) and gp552 (U2) (Figure 36). Proteins in the bands near the interface and the two upper bands formed in the T2 lane were identified by mass spectrometry. We observe no significant difference between C2 and C3, or between T2 and T3. The mass spectrometry data is listed in Table 7.

Figure 36 121Q protein gel bands for mass spectrometry



Names	Samples
Phage	Phage
C2	Capsid
C3	Capsid with trypsin treatment
T2	Tail
T3	Tail with trypsin treatment
U1	Separated band 1
U2	Separated band 2

Table 7 Spectrum counts of structural proteins in mass spectrometry

ID	MW (kDa)	Phage	C2	C3	U1	U2	T2	T3	C/T ratio	Protein
567	19	9	5	2	0	0	0	0	∞	
584	42	8	11	1	0	0	0	0	∞	
5	48	14	3	1	0	0	0	0	∞	
38	13	3	3	1	0	0	0	0	∞	
6	39	2	2	1	0	0	0	0	∞	
493	12	0	2	0	0	0	0	0	∞	
24	35	5	0	0	0	0	0	0	∞	
25	8	3	0	0	0	0	0	0	∞	
30	21	0	2	0	0	0	0	0	∞	
45	21	66	42	80	9	1	5	15	6.93	
51	29	75	7	3	0	0	0	2	6.38	
583	23	20	3	8	0	0	1	1	5.96	Protease
50	60	43	23	36	0	0	0	18	3.88	
452	14	32	30	47	5	1	11	17	3.11	
147	43	16	18	31	1	0	6	12	3.08	
600	21	449	389	513	159	62	103	236	3.07	Dimeric surface protein
193	55	57	37	44	0	0	4	28	3.00	
540	57	36	13	9	0	0	1	8	3.00	
582	31	16	15	12	0	0	2	9	2.96	
599	24	525	512	577	252	60	135	314	2.83	Dimeric surface protein
191	15	27	19	12	0	0	9	4	2.76	
580	65	39	31	36	1	1	13	17	2.56	Portal
27	23	28	21	22	2	1	6	14	2.52	
39	12	1	2	0	0	0	1	0	2.45	
28	12	3	1	1	0	0	0	1	2.43	
585	42	601	621	762	264	82	281	379	2.40	MCP
49	21	12	3	3	0	0	1	2	2.34	
499	17	28	18	21	1	0	6	14	2.27	
538	21	30	26	28	3	0	10	18	2.24	
63	26	11	6	7	0	0	2	5	2.16	
541	36	4	1	3	1	0	1	1	2.16	
542	14	7	6	9	0	0	3	5	2.13	
528	30	15	8	16	1	0	7	7	1.89	Neck?
62	21	16	10	8	0	0	2	10	1.81	
233	24	7	2	4	0	0	2	2	1.65	
539	20	11	7	6	0	0	3	7	1.54	

237	41	10	4	5	0	0	4	4	1.27	
15	13	3	3	0	0	0	3	0	1.22	
235	41	2	2	4	0	0	3	3	1.10	
18	37	5	5	8	0	0	9	6	0.96	
226	39	1	3	3	0	0	5	3	0.85	
229	35	0	1	2	0	0	2	2	0.83	
234	43	3	5	14	0	0	12	14	0.79	
581	9	3	1	1	0	0	1	2	0.78	
236	39	3	6	13	0	0	13	16	0.72	
238	41	1	5	7	0	0	10	9	0.71	
228	30	10	4	3	0	0	6	7	0.63	
556	19	3	1	1	0	0	1	3	0.59	
31	21	4	5	12	1	0	15	17	0.58	
481	21	3	3	4	0	0	7	8	0.53	
231	40	2	2	6	0	0	8	12	0.44	
566	15	1	0	2	0	1	1	4	0.42	
40	25	6	2	2	0	0	0	12	0.41	
232	41	0	3	7	0	0	15	14	0.38	
561	57	11	3	4	0	0	11	10	0.38	
560	52	8	6	5	0	0	19	16	0.36	Tail sheath stabilizer and completion
557	22	6	2	6	0	0	10	15	0.35	
533	25	7	4	7	2	4	19	16	0.35	
534	29	8	3	5	1	1	15	12	0.33	
555	41	72	18	30	2	3	98	69	0.32	
553	31	5	4	6	0	5	15	22	0.31	
558	40	35	9	19	0	2	52	49	0.30	
535	40	26	7	13	1	0	37	37	0.30	
554	78	20	16	21	2	4	82	59	0.29	
482	96	1	12	23	5	12	70	64	0.29	
54	90	5	9	20	0	3	65	45	0.29	
564	131	6	18	41	46	77	108	119	0.28	Baseplate wedge
579	80	49	13	22	0	5	77	72	0.26	
480	99	0	8	11	14	19	41	44	0.25	
531	97	39	97	142	37	55	565	586	0.23	Sheath
570	12	5	1	2	0	0	9	6	0.22	
569	85	23	7	7	0	24	37	37	0.22	
550	70	7	14	28	8	3	120	98	0.21	
488	5	4	1	1	0	0	6	5	0.21	
489	19	7	2	2	0	0	11	13	0.19	
568	96	2	5	10	6	7	44	44	0.19	Baseplate wedge

589	57	3	9	25	1	0	122	117	0.15	Tail fiber
552	180	10	26	30	76	1608	298	135	0.14	Hair
563	382	2	1	7	480	277	40	61	0.08	Tail fiber?
571	28	8	0	2	1	0	14	17	0.07	
490	105	2	1	1	0	2	29	26	0.04	Tail fiber
106	8	0	0	0	0	0	9	0	0.00	
514	12	0	0	0	0	0	2	0	0.00	
Total	-	2650	2250	2888	1382	2320	2755	3076	-	-

Notes:

The ratio of the normalized spectrum count of a protein from the capsid sample to the count from the tail sample (or C/T ratio) for a protein is calculated as following:

Normalized spectrum count (N-count) = the count of the protein / the total count of all 121Q proteins from the sample;

C/T ratio = (N-count from C2 + N-count from C3) / (N-count from T2 + N-count from T3).

The rows are sorted in descending order based on C/T ratio. Proteins with total spectrum counts less than 3 from the tail and capsid samples are not listed in (Table 3).

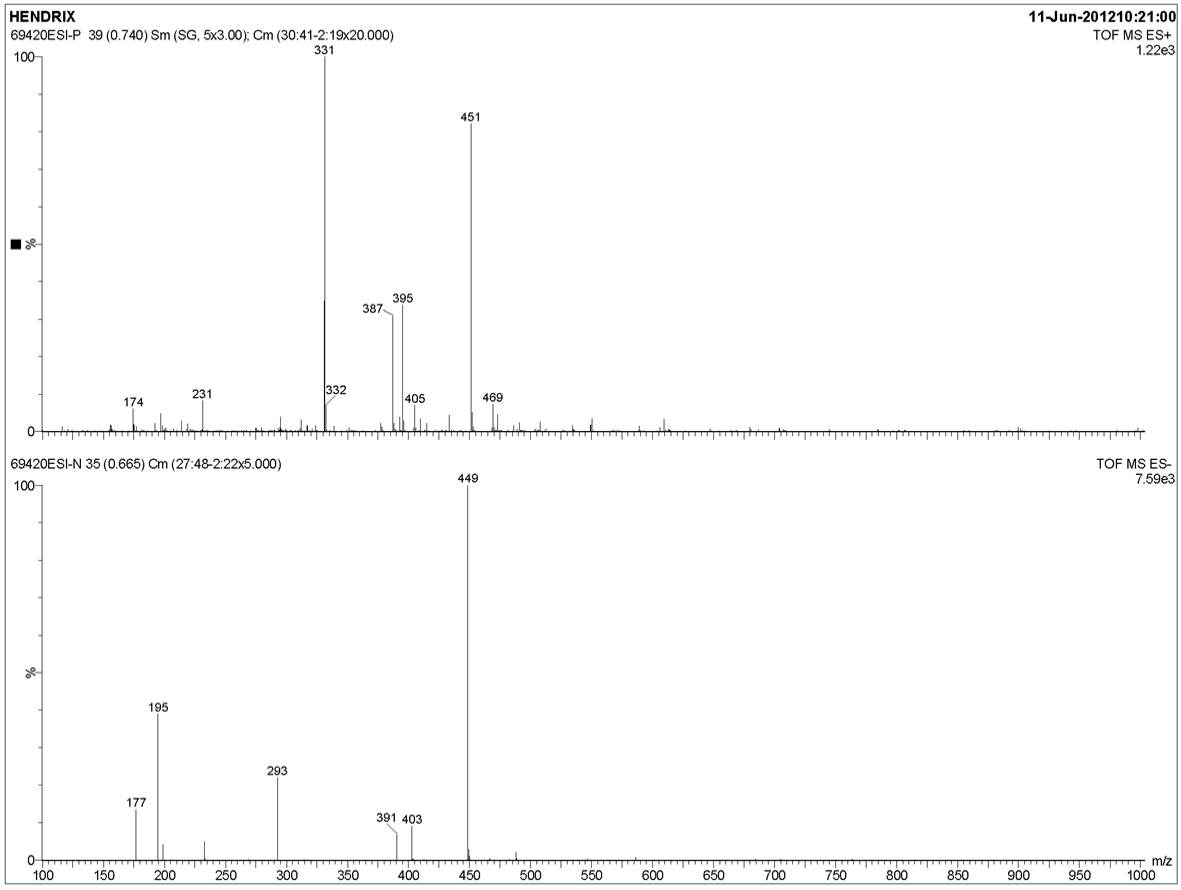
APPENDIX C

MASS SPECTROMETRY OF THE MODIFIED CYTIDINE IN PAU DNA

Modified cytidine in PAU DNA was isolated by HPLC. The purified modified cytidine was dried in a SpeedVac in Buffer B and sent for time-of-flight mass spectrometry with electrospray ionization. But later we found that the modified cytidine could not be recovered after SpeedVac. Two major species with molecular weight at 450 and 330 g/mol were detected, much higher than the cytidine (253 g/mol). We were unable to find any modified cytidine to match either mass after checking existing DNA/RNA modification libraries. We thought that oxidation might be the reason why the modified base could not be recovered. Later we reduced the volume of the modified cytidine in water with N_2 flow to about 5 uL and sent the sample to the mass spectrometry facility again. The result was received about 3 weeks later, but the spectra were very noisy. No significant difference was observed between the sample spectra and the control (data not shown).

Figure 37 Time-of-flight mass spectrometry spectra with electrospray ionization.

Upper: ES+; Lower: ES-.



BIBLIOGRAPHY

Abbasifar, R., M. W. Griffiths, P. M. Sabour, H. W. Ackermann, K. Vandersteegen, R. Lavigne, J. P. Noben, A. Alanis Villa, A. Abbasifar, J. H. Nash and A. M. Kropinski (2014). "Supersize me: Cronobacter sakazakii phage GAP32." Virology **460-461**: 138-146.

Ackermann, H. W. (2001). "Frequency of morphological phage descriptions in the year 2000. Brief review." Arch Virol **146**(5): 843-857.

Ackermann, H. W., P. Auclair, S. Basavarajappa, H. P. Konjin and C. Savanurmath (1994). "Bacteriophages from Bombyx mori." Arch Virol **137**(1-2): 185-190.

Ackermann, H. W. and T. M. Nguyen (1983). "Sewage coliphages studied by electron microscopy." Appl Environ Microbiol **45**(3): 1049-1059.

Adrian, M., J. Dubochet, J. Lepault and A. W. McDowall (1984). "Cryo-electron microscopy of viruses." Nature **308**(5954): 32-36.

Bachrach, U. and L. Benchetrit (1974). "Studies on phage internal proteins. 3. Specific binding of T4 internal proteins to T4 DNA." Virology **59**(2): 443-454.

Baker, T. S., N. H. Olson and S. D. Fuller (1999). "Adding the third dimension to virus life cycles: three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs." Microbiol Mol Biol Rev **63**(4): 862-922, table of contents.

Bauer, D. W. and A. Evilevitch (2015). "Influence of Internal DNA Pressure on Stability and Infectivity of Phage lambda." J Mol Biol **427**(20): 3189-3200.

Bauer, D. W., D. Li, J. Huffman, F. L. Homa, K. Wilson, J. C. Leavitt, S. R. Casjens, J. Baines and A. Evilevitch (2015). "Exploring the Balance between DNA Pressure and Capsid Stability in Herpesviruses and Phages." J Virol **89**(18): 9288-9298.

Bennion, B. J. and V. Daggett (2003). "The molecular basis for the chemical denaturation of proteins by urea." Proc Natl Acad Sci U S A **100**(9): 5142-5147.

Bhella, D., F. J. Rixon and D. J. Dargan (2000). "Cryomicroscopy of human cytomegalovirus virions reveals more densely packed genomic DNA than in herpes simplex virus type 1." J Mol Biol **295**(2): 155-161.

Bijlenga, R. K., T. Ishii and A. Tsugita (1978). "Complete primary structure of the small outer capsid (soc) protein of bacteriophage T4." J Mol Biol **120**(2): 249-263.

Black, L. W. (1974). "Bacteriophage T4 internal protein mutants: isolation and properties." Virology **60**(1): 166-179.

Black, L. W. (1989). "DNA packaging in dsDNA bacteriophages." Annu Rev Microbiol **43**: 267-292.

Brenner, D. J., Krieg, Noel R., Staley, James T. (Eds.) (2005). The Proteobacteria, Springer.

Brenner, S., M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S. R. Williams, K. Moon, T. Burcham, M. Pallas, R. B. DuBridge, J. Kirchner, K. Fearon, J. Mao and K.

Corcoran (2000). "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays." Nat Biotechnol **18**(6): 630-634.

Casjens, S. and M. Hayden (1988). "Analysis in vivo of the bacteriophage P22 headful nuclease." J Mol Biol **199**(3): 467-474.

Casjens, S. and R. Hendrix (1988). dsDNA Phage Assembly. The Bacteriophages. R. Calendar. New York, Plenum Press.

Casjens, S. and J. King (1975). "Virus assembly." Annu Rev Biochem **44**: 555-611.

Casjens, S. R. and E. B. Gilcrease (2009). "Determining DNA packaging strategy by analysis of the termini of the chromosomes in tailed-bacteriophage virions." Methods Mol Biol **502**: 91-111.

Caspar, D. L. and A. Klug (1962). "Physical principles in the construction of regular viruses." Cold Spring Harb Symp Quant Biol **27**: 1-24.

Chow, L. T. and A. I. Bukhari (1978). "Heteroduplex electron microscopy of phage Mu mutants containing IS1 insertions and chloramphenicol resistance transposons." Gene **3**(4): 333-346.

Conley, M. P. and W. B. Wood (1975). "Bacteriophage T4 whiskers: a rudimentary environment-sensing device." Proc Natl Acad Sci U S A **72**(9): 3701-3705.

Conway, J. F., N. Cheng, A. Zlotnick, P. T. Wingfield, S. J. Stahl and A. C. Steven (1997). "Visualization of a 4-helix bundle in the hepatitis B virus capsid by cryo-electron microscopy." Nature **386**(6620): 91-94.

Conway, J. F. and A. C. Steven (1999). "Methods for reconstructing density maps of "single" particles from cryoelectron micrographs to subnanometer resolution." J Struct Biol **128**(1): 106-118.

Conway, J. F., W. R. Wikoff, N. Cheng, R. L. Duda, R. W. Hendrix, J. E. Johnson and A. C. Steven (2001). "Virus maturation involving large subunit rotations and local refolding." Science **292**(5517): 744-748.

Cornelissen, A., S. C. Hardies, O. V. Shaburova, V. N. Krylov, W. Mattheus, A. M. Kropinski and R. Lavigne (2012). "Complete genome sequence of the giant virus OBP and comparative genome analysis of the diverse PhiKZ-related phages." J Virol **86**(3): 1844-1852.

Cregg, J. M. and C. R. Stewart (1978). "Terminal redundancy of "high frequency of recombination" markers of Bacillus subtilis phage SPO1." Virology **86**(2): 530-541.

Crick, F. H. and J. D. Watson (1956). "Structure of small viruses." Nature **177**(4506): 473-475.

d'Herelle, F. (1917). "Sur un microbe invisible antagoniste des bacilles dysenteriques." Compt Rend Acad Sci **165**: 373.

Duda, R. L., R. W. Hendrix, W. M. Huang and J. F. Conway (2006). "Shared architecture of bacteriophage SPO1 and herpesvirus capsids." Curr Biol **16**(1): R11-13.

Duda, R. L., K. Martincic and R. W. Hendrix (1995). "Genetic basis of bacteriophage HK97 prohead assembly." J Mol Biol **247**(4): 636-647.

Earnshaw, W. C. and S. C. Harrison (1977). "DNA arrangement in isometric phage heads." Nature **268**(5621): 598-602.

Edman, P. (1950). "Method for Determination of the Amino Acid Sequence in Peptides." Acta Chemica Scandinavica **4**(2): 283-293.

Effantin, G., P. Boulanger, E. Neumann, L. Letellier and J. F. Conway (2006). "Bacteriophage T5 structure reveals similarities with HK97 and T4 suggesting evolutionary relationships." J Mol Biol **361**(5): 993-1002.

Effantin, G., R. Hamasaki, T. Kawasaki, M. Bacia, C. Moriscot, W. Weissenhorn, T. Yamada and G. Schoehn (2013). "Cryo-electron microscopy three-dimensional structure of the jumbo phage PhiRSL1 infecting the phytopathogen Ralstonia solanacearum." Structure **21**(2): 298-305.

Fauquet, C. M. and D. Fargette (2005). "International Committee on Taxonomy of Viruses and the 3,142 unassigned species." Virol J **2**: 64.

Feiss, M. and W. Widner (1982). "Bacteriophage lambda DNA packaging: scanning for the terminal cohesive end site during packaging." Proc Natl Acad Sci U S A **79**(11): 3498-3502.

Fokine, A., P. R. Chipman, P. G. Leiman, V. V. Mesyanzhinov, V. B. Rao and M. G. Rossmann (2004). "Molecular architecture of the prolate head of bacteriophage T4." Proc Natl Acad Sci U S A **101**(16): 6003-6008.

Fokine, A., V. A. Kostyuchenko, A. V. Efimov, L. P. Kurochkina, N. N. Sykilinda, J. Robben, G. Volckaert, A. Hoenger, P. R. Chipman, A. J. Battisti, M. G. Rossmann and V. V. Mesyanzhinov (2005). "A three-dimensional cryo-electron microscopy structure of the bacteriophage phiKZ head." J Mol Biol **352**(1): 117-124.

Fokine, A., P. G. Leiman, M. M. Shneider, B. Ahvazi, K. M. Boeshans, A. C. Steven, L. W. Black, V. V. Mesyanzhinov and M. G. Rossmann (2005). "Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to a common ancestry." Proc Natl Acad Sci U S A **102**(20): 7163-7168.

Folch, B., Y. Dehouck and M. Rooman (2010). "Thermo- and mesostabilizing protein interactions identified by temperature-dependent statistical potentials." Biophys J **98**(4): 667-677.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan and J. Postlethwait (1999). "Preservation of duplicate genes by complementary, degenerative mutations." Genetics **151**(4): 1531-1545.

Gipson, P., M. L. Baker, D. Raytcheva, C. Haase-Pettingell, J. Piret, J. A. King and W. Chiu (2014). "Protruding knob-like proteins violate local symmetries in an icosahedral marine virus." Nat Commun **5**: 4278.

Glupczynski, Y., W. Hansen, M. Dratwa, C. Tielemans, R. Wens, F. Collart and E. Yourassowsky (1984). "Pseudomonas paucimobilis peritonitis in patients treated by peritoneal dialysis." J Clin Microbiol **20**(6): 1225-1226.

Goff, C. G. (1979). "Bacteriophage T4 alt gene maps between genes 30 and 54." J Virol **29**(3): 1232-1234.

Gommers-Ampt, J. H. and P. Borst (1995). "Hypermodified bases in DNA." FASEB J **9**(11): 1034-1042.

Guo, F., Z. Liu, P. A. Fang, Q. Zhang, E. T. Wright, W. Wu, C. Zhang, F. Vago, Y. Ren, J. Jakana, W. Chiu, P. Serwer and W. Jiang (2014). "Capsid expansion mechanism of bacteriophage T7 revealed by multistate atomic models derived from cryo-EM reconstructions." Proc Natl Acad Sci U S A **111**(43): E4606-4614.

Harpaz, Y., M. Gerstein and C. Chothia (1994). "Volume changes on protein folding." Structure **2**(7): 641-649.

Hendrix, R. W. (1999). "Evolution: the long evolutionary reach of viruses." Curr Biol **9**(24): R914-917.

Hendrix, R. W. (2009). "Jumbo bacteriophages." Curr Top Microbiol Immunol **328**: 229-240.

Hendrix, R. W., M. C. Smith, R. N. Burns, M. E. Ford and G. F. Hatfull (1999). "Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage." Proc Natl Acad Sci U S A **96**(5): 2192-2197.

Hershey, A. D. and M. Chase (1952). "Independent functions of viral protein and nucleic acid in growth of bacteriophage." J Gen Physiol **36**(1): 39-56.

Heymann, J. B. (2001). "Bsoft: image and molecular processing in electron microscopy." J Struct Biol **133**(2-3): 156-169.

Hua, J. (2010). Regulation of capsid sizes of large tailed bacteriophages. MS, University of Pittsburgh.

Huang, L. H., C. M. Farnet, K. C. Ehrlich and M. Ehrlich (1982). "Digestion of highly modified bacteriophage DNA by restriction endonucleases." Nucleic Acids Res **10**(5): 1579-1591.

Humphreys, G. O. and T. A. Trautner (1981). "Maturation of bacteriophage SPPI DNA: limited precision in the sizing of mature bacteriophage genomes." J Virol **37**(2): 832-835.

Iida, S., M. B. Streiff, T. A. Bickle and W. Arber (1987). "Two DNA antirestriction systems of bacteriophage P1, darA, and darB: characterization of darA- phages." Virology **157**(1): 156-166.

Jiang, W., M. L. Baker, J. Jakana, P. R. Weigele, J. King and W. Chiu (2008). "Backbone structure of the infectious epsilon15 virus capsid revealed by electron cryomicroscopy." Nature **451**(7182): 1130-1134.

Jonczyk, E., M. Klak, R. Miedzybrodzki and A. Gorski (2011). "The influence of external factors on bacteriophages--review." Folia Microbiol (Praha) **56**(3): 191-200.

Kaliman, A. V., M. A. Khasanova, V. M. Kryukov, V. I. Tanyashin and A. A. Bayev (1990). "The nucleotide sequence of the region of bacteriophage T4 inh(lip)-hoc genes." Nucleic Acids Res **18**(14): 4277.

Kato, H. and C. Baschong (1997). "Isolation of a gp20-complex and its role in in vitro assembly of both prohead and core of bacteriophage T4." Virology **227**(2): 400-408.

Keen, E. C. (2012). "Phage therapy: concept to cure." Front Microbiol **3**: 238.

Keller, B., E. Kellenberger, T. A. Bickle and A. Tsugita (1985). "Determination of the cleavage site of the phage T4 prohead protease in gene product 68. Influence of protein secondary structure on cleavage specificity." J Mol Biol **186**(3): 665-667.

Kelley, L. A. and M. J. Sternberg (2009). "Protein structure prediction on the Web: a case study using the Phyre server." Nat Protoc **4**(3): 363-371.

Kelly, T. J., Jr. and C. A. Thomas, Jr. (1969). "An intermediate in the replication of bacteriophage T7 DNA molecules." J Mol Biol **44**(3): 459-475.

Kim, M. S., S. S. Hong, K. Park and H. Myung (2013). "Genomic analysis of bacteriophage PBECO4 infecting Escherichia coli O157:H7." Arch Virol **158**(11): 2399-2403.

Krylov, V. N., T. A. Smirnova, B. A. Rebentish and I. B. Minenkova (1978). "[Stucture of PhiKZ bacteriophage particles]." Vopr Virusol(5): 568-571.

Kuo, K. C., R. A. McCune, C. W. Gehrke, R. Midgett and M. Ehrlich (1980). "Quantitative reversed-phase high performance liquid chromatographic determination of major and modified deoxyribonucleosides in DNA." Nucleic Acids Res **8**(20): 4763-4776.

Kurtz, M. B. and S. P. Champe (1977). "Precursors of the T4 internal peptides." J Virol **22**(2): 412-419.

Lander, G. C., A. C. Baudoux, F. Azam, C. S. Potter, B. Carragher and J. E. Johnson (2012). "Capsomer dynamics and stabilization in the T = 12 marine bacteriophage SIO-2 and its procapsid studied by CryoEM." Structure **20**(3): 498-503.

Lander, G. C., A. Evilevitch, M. Jeembaeva, C. S. Potter, B. Carragher and J. E. Johnson (2008). "Bacteriophage lambda stabilization by auxiliary protein gpD: timing, location, and mechanism of attachment determined by cryo-EM." Structure **16**(9): 1399-1406.

Lander, G. C., L. Tang, S. R. Casjens, E. B. Gilcrease, P. Prevelige, A. Poliakov, C. S. Potter, B. Carragher and J. E. Johnson (2006). "The structure of an infectious P22 virion shows the signal for headful DNA packaging." Science **312**(5781): 1791-1795.

Lederberg, J. (1952). "Cell genetics and hereditary symbiosis." Physiol Rev **32**(4): 403-430.

Leiman, P. G., F. Arisaka, M. J. van Raaij, V. A. Kostyuchenko, A. A. Aksyuk, S. Kanamaru and M. G. Rossmann (2010). "Morphogenesis of the T4 tail and tail fibers." Virol J **7**: 355.

Lennox, E. S. (1955). "Transduction of linked genetic characters of the host by bacteriophage P1." Virology **1**(2): 190-206.

Lim, W. K., J. Rosgen and S. W. Englander (2009). "Urea, but not guanidinium, destabilizes proteins by forming hydrogen bonds to the peptide group." Proc Natl Acad Sci U S A **106**(8): 2595-2600.

Lobocka, M. B., D. J. Rose, G. Plunkett, 3rd, M. Rusin, A. Samojedny, H. Lehnherr, M. B. Yarmolinsky and F. R. Blattner (2004). "Genome of bacteriophage P1." J Bacteriol **186**(21): 7032-7068.

Locker, C. R., S. D. Fuller and S. C. Harvey (2007). "DNA organization and thermodynamics during viral packing." Biophys J **93**(8): 2861-2869.

Masai, E., Y. Katayama, S. Nishikawa and M. Fukuda (1999). "Characterization of Sphingomonas paucimobilis SYK-6 genes involved in degradation of lignin-related compounds." J Ind Microbiol Biotechnol **23**(4-5): 364-373.

Mesyanzhinov, V. V., J. Robben, B. Grymonprez, V. A. Kostyuchenko, M. V. Bourkaltseva, N. N. Sykilinda, V. N. Krylov and G. Volckaert (2002). "The genome of bacteriophage phiKZ of Pseudomonas aeruginosa." J Mol Biol **317**(1): 1-19.

Miller, E. S., J. F. Heidelberg, J. A. Eisen, W. C. Nelson, A. S. Durkin, A. Ciecko, T. V. Feldblyum, O. White, I. T. Paulsen, W. C. Nierman, J. Lee, B. Szczypinski and C. M. Fraser (2003). "Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage." J Bacteriol **185**(17): 5220-5233.

Monera, O. D., C. M. Kay and R. S. Hodges (1994). "Protein denaturation with guanidine hydrochloride or urea provides a different estimate of stability depending on the contributions of electrostatic interactions." Protein Sci **3**(11): 1984-1991.

Morais, M. C., K. H. Choi, J. S. Koti, P. R. Chipman, D. L. Anderson and M. G. Rossmann (2005). "Conservation of the capsid structure in tailed dsDNA bacteriophages: the pseudoatomic structure of phi29." Mol Cell **18**(2): 149-159.

Nirenberg, M. W. and J. H. Matthaei (1961). "The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides." Proc Natl Acad Sci U S A **47**: 1588-1602.

Ohno, S. (1970). Evolution by Gene Duplication, Springer.

Oppenheim, A. B., A. J. Rattray, M. Bubunenko, L. C. Thomason and D. L. Court (2004). "In vivo recombineering of bacteriophage lambda by PCR fragments and single-strand oligonucleotides." Virology **319**(2): 185-189.

Parent, K. N., E. B. Gilcrease, S. R. Casjens and T. S. Baker (2012). "Structural evolution of the P22-like phages: comparison of Sf6 and P22 procapsid and virion architectures." Virology **427**(2): 177-188.

Pettersen, E. F., T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng and T. E. Ferrin (2004). "UCSF Chimera--a visualization system for exploratory research and analysis." J Comput Chem **25**(13): 1605-1612.

Purohit, P. K., J. Kondev and R. Phillips (2003). "Mechanics of DNA packaging in viruses." Proc Natl Acad Sci U S A **100**(6): 3173-3178.

Rau, D. C., B. Lee and V. A. Parsegian (1984). "Measurement of the repulsive force between polyelectrolyte molecules in ionic solution: hydration forces between parallel DNA double helices." Proc Natl Acad Sci U S A **81**(9): 2621-2625.

Ronaghi, M., S. Karamohamed, B. Pettersson, M. Uhlen and P. Nyren (1996). "Real-time DNA sequencing using detection of pyrophosphate release." Anal Biochem **242**(1): 84-89.

Ryan, M. P. and C. C. Adley (2010). "Sphingomonas paucimobilis: a persistent Gram-negative nosocomial infectious organism." Journal of Hospital Infection **75**(3): 153-157.

Rydman, P. S., J. K. Bamford and D. H. Bamford (2001). "A minor capsid protein P30 is essential for bacteriophage PRD1 capsid assembly." J Mol Biol **313**(4): 785-795.

Salas, M., C. Vasquez, E. Mendez and E. Vinuela (1972). "Head fibers of bacteriophage phi 29." Virology **50**(1): 180-188.

Schwartz, D. C. and C. R. Cantor (1984). "Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis." Cell **37**(1): 67-75.

Schwarzer, D., F. F. Buettner, C. Browning, S. Nazarov, W. Rabsch, A. Bethe, A. Oberbeck, V. D. Bowman, K. Stummeyer, M. Muhlenhoff, P. G. Leiman and R. Gerardy-Schahn (2012). "A multivalent adsorption apparatus explains the broad host range of phage phi92: a comprehensive genomic and structural analysis." J Virol **86**(19): 10384-10398.

Shaikh, T. R., H. Gao, W. T. Baxter, F. J. Asturias, N. Boisset, A. Leith and J. Frank (2008). "SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs." Nat Protoc **3**(12): 1941-1974.

Shen, P. S., M. J. Domek, E. Sanz-Garcia, A. Makaju, R. M. Taylor, R. Hoggan, M. D. Culumber, C. J. Oberg, D. P. Breakwell, J. T. Prince and D. M. Belnap (2012). "Sequence and structural characterization of great salt lake bacteriophage CW02, a member of the T7-like supergroup." J Virol **86**(15): 7907-7917.

Simoliunas, E., L. Kaliniene, L. Truncaite, A. Zajanckauskaite, J. Staniulis, A. Kaupinis, M. Ger, M. Valius and R. Meskys (2013). "Klebsiella phage vB_KleM-RaK2 - a giant singleton virus of the family Myoviridae." PLoS One **8**(4): e60717.

Smith, D. E., S. J. Tans, S. B. Smith, S. Grimes, D. L. Anderson and C. Bustamante (2001). "The bacteriophage straight phi29 portal motor can package DNA against a large internal force." Nature **413**(6857): 748-752.

Soding, J. (2005). "Protein homology detection by HMM-HMM comparison." Bioinformatics **21**(7): 951-960.

Spanova, A. (1992). "Comparison of permuted region lengths in the genomes of related Salmonella typhimurium phages P22 and L." Folia Microbiol (Praha) **37**(3): 188-192.

Sternberg, N. and R. Weisberg (1977). "Packaging of coliphage lambda DNA. II. The role of the gene D protein." J Mol Biol **117**(3): 733-759.

Streisinger, G., R. S. Edgar and G. H. Denhardt (1964). "Chromosome Structure in Phage T4. I. Circularity of the Linkage Map." Proc Natl Acad Sci U S A **51**: 775-779.

Sugimoto, K., T. Okazaki and R. Okazaki (1968). "Mechanism of DNA chain growth, II. Accumulation of newly synthesized short chains in E. coli infected with ligase-defective T4 phages." Proc Natl Acad Sci U S A **60**(4): 1356-1362.

Suhanovsky, M. M. and C. M. Teschke (2015). "Nature's favorite building block: Deciphering folding and capsid assembly of proteins with the HK97-fold." Virology **479-480**: 487-497.

Suttle, C. A. (2005). "Viruses in the sea." Nature **437**(7057): 356-361.

Tada, Y. and T. Inoue (2000). "Use of oligotrophic bacteria for the biological monitoring of heavy metals." Journal of Applied Microbiology **88**(1): 154-160.

Takahashi, I. (1961). "Genetic transduction in Bacillus subtilis." Biochem Biophys Res Commun **5**: 171-175.

Tang, L., E. B. Gilcrease, S. R. Casjens and J. E. Johnson (2006). "Highly discriminatory binding of capsid-cementing proteins in bacteriophage L." Structure **14**(5): 837-845.

Tao, Y., N. H. Olson, W. Xu, D. L. Anderson, M. G. Rossmann and T. S. Baker (1998). "Assembly of a tailed bacterial virus and its genome release studied in three dimensions." Cell **95**(3): 431-437.

Tavares, P., R. Lurz, A. Stiege, B. Ruckert and T. A. Trautner (1996). "Sequential headful packaging and fate of the cleaved DNA ends in bacteriophage SPP1." J Mol Biol **264**(5): 954-967.

Thomas, J. A., S. T. Weintraub, W. Wu, D. C. Winkler, N. Cheng, A. C. Steven and L. W. Black (2012). "Extensive proteolysis of head and inner body proteins by a morphogenetic protease in the giant Pseudomonas aeruginosa phage phiKZ." Mol Microbiol **84**(2): 324-339.

Toropova, K., J. B. Huffman, F. L. Homa and J. F. Conway (2011). "The herpes simplex virus 1 UL17 protein is the second constituent of the capsid vertex-specific component required for DNA packaging and retention." J Virol **85**(15): 7513-7522.

Traub, F., B. Keller, A. Kuhn and M. Maeder (1984). "Isolation of the prohead core of bacteriophage T4 after cross-linking and determination of protein composition." J Virol **49**(3): 902-908.

Twort, F. W. (1915). "An investigation on thenature of ultra-microscopic viruses." Lancet **II**: 1241.

van Heel, M., B. Gowen, R. Matadeen, E. V. Orlova, R. Finn, T. Pape, D. Cohen, H. Stark, R. Schmidt, M. Schatz and A. Patwardhan (2000). "Single-particle electron cryo-microscopy: towards atomic resolution." Q Rev Biophys **33**(4): 307-369.

Walker, D. H., Jr. and T. F. Anderson (1970). "Morphological variants of coliphage P1." J Virol **5**(6): 765-782.

Warren, R. A. (1980). "Modified bases in bacteriophage DNAs." Annu Rev Microbiol **34**: 137-158.

Weigele, P. R., W. H. Pope, M. L. Pedulla, J. M. Houtz, A. L. Smith, J. F. Conway, J. King, G. F. Hatfull, J. G. Lawrence and R. W. Hendrix (2007). "Genomic and structural analysis of Syn9, a cyanophage infecting marine Prochlorococcus and Synechococcus." Environ Microbiol **9**(7): 1675-1695.

Wikoff, W. R., L. Liljas, R. L. Duda, H. Tsuruta, R. W. Hendrix and J. E. Johnson (2000). "Topologically linked protein rings in the bacteriophage HK97 capsid." Science **289**(5487): 2129-2133.

Wu, W., J. A. Thomas, N. Cheng, L. W. Black and A. C. Steven (2012). "Bubblegrams reveal the inner body of bacteriophage phiKZ." Science **335**(6065): 182.

Wyatt, G. R. (1953). "The quantitative composition of deoxypentose nucleic acids as related to the newly proposed structure." Cold Spring Harb Symp Quant Biol **18**: 133-134.

Wyatt, G. R. and S. S. Cohen (1953). "The base of the desoxyribonucleic acids of T2, T4, and T6 bacteriophages." Ann Inst Pasteur (Paris) **84**(1): 143-146.

Yan, X., R. S. Sinkovits and T. S. Baker (2007). "AUTO3DEM--an automated and high throughput program for image reconstruction of icosahedral particles." J Struct Biol **157**(1): 73-82.

Yang, J., R. Yan, A. Roy, D. Xu, J. Poisson and Y. Zhang (2015). "The I-TASSER Suite: protein structure and function prediction." Nat Methods **12**(1): 7-8.